

Original Article

# Multimodal Feature-Based Deep Learning Framework for Person Re-Identification: Enhancing Models with InceptionNet Representation

Badireddygar Anurag Reddy<sup>1</sup>, Danvir Mandal<sup>2</sup>, Bhaveshkumar C. Dharmani<sup>3</sup>

<sup>1</sup>Department of Electronics and Communication Engineering, Lovely Professional University, Punjab, India.

<sup>2</sup>Department of Interdisciplinary Courses in Engineering (DICE), Chitkara University, Punjab, India.

<sup>3</sup>SEEE, Lovely Professional University, Punjab, India.

<sup>1</sup>Corresponding Author : [anuragreddy402@gmail.com](mailto:anuragreddy402@gmail.com)

Received: 11 March 2025

Revised: 24 May 2025

Accepted: 19 June 2025

Published: 30 July 2025

**Abstract** - In security, surveillance, and identity verification systems, person re-identification, or Re-ID, has become a vital task in the field of computer vision since its introduction. Conventional methods sometimes lead one to run across issues, including changing viewpoints, shadows, and various lighting conditions. Recent advances in deep learning, allowing the use of multimodal data and robust feature extraction techniques, have produced interesting ideas. In this work, a deep learning-based method for person re-identification is investigated using the Deep Multimodal Inception Network Representation Learning (DMIRL) framework. Review of pre-existing Re-ID algorithms on widely used datasets, including DukeMTMC-reID and Market-1501, comes first in the process. Various approaches to data preparation are used to improve the datasets. These methods consist of augmentation, image normalisation, and multimodal feature extraction. An advanced InceptionNet architecture capable of learning complementary features from multimodal inputs is used in the proposed DMIRL model. Among these inputs are optical, infrared, and skeletal ones. Experimental analyses reveal that DMIRL successfully addresses pose variations and partial occlusions. The proposed method achieved 89.5% accuracy on MSMT17, 92.4% on Market1501, 91.1% on DukeMTMC, and 85.3% on CUHK03-NP. Mean Average Precision (mAP) reached 87.6% on Market1501. Computational efficiency ranged from 0.30s to 0.40s. Cross-modality (RGB to IR) showed a slight decline, maintaining 85.0% accuracy on MSMT17.

**Keywords** - Deep learning, InceptionNet, Multimodal features, Person re-identification, Representation learning.

## 1. Introduction

Person Re-Identification (ReID) has become rather important in the field of computer vision. For security applications and surveillance systems, where the objective is exactly to identify a person across several camera views, this is particularly true. ReID poses problems regarding the notable variations in appearance resulting from camera angles, lighting conditions, poses, and occlusions. Deep learning, attention mechanisms, and multi-modal fusion are among the developments that ReID methods have seen. These developments have been used to improve identification accuracy in demanding environments. Conventional RGB-only models fail in this regard; hence, the integration of multimodal data, such as Infrared (IR) and RGB images, further enriches the ReID process by allowing strong recognition under a range of lighting conditions and at different times of the day [1-3]. Although deep learning-based models have succeeded in personal re-identification, several issues still need to be addressed to provide more pertinent and effective

systems. Particularly for big datasets like MSMT17 and Market1501, one of the main challenges is the high computational cost associated with training and inference. Each one of these datasets contains thousands of identities, each of which is subject to a different set of environmental conditions; hence, the model has a challenging optimising task [4]. Moreover, a big difficulty is performance in several modes. Apart from making the re-identification process more difficult, the variations between RGB and IR images complicate the establishment of correct correspondences between images of the same person taken using several modalities [5]. Memory consumption is another main obstacle, particularly in real-time applications that typically call for questions regarding the hardware constraints and model size [6]. Moreover, even if this is required for deployment in settings with limited resources, energy economy is generally neglected in favour of better performance [7]. The main challenge this work tackles is the enhancement of person Re-Identification (ReID) in cross-



modality environments, including the application of both RGB and Infrared (IR) modalities. The methods now in use find it difficult to achieve high accuracy over all ranges, given the great variations in data sources, lighting conditions, and camera types. Using multi-modal data could provide a solution; yet, if one is to properly combine several modalities, one must overcome the challenges of alignment, feature extraction, and representation learning.

Moreover, the difficulty of present models' implementation in the real world, particularly on devices with limited resources, is due to their high computational demands and energy inefficiencies found in training and inference [8-12]. The goal of this work is to develop a new ReID framework that is able to concurrently address computational efficiency, memory consumption, and energy economy by efficiently combining multi-modal features (RGB and IR images).

The suggested method aims to provide a more solid, scalable, and effective solution to person re-identification across several camera kinds and lighting conditions. Modern deep learning techniques and well-crafted networks help to accomplish this.

The main objective of this work is to develop a cross-modality ReID method addressing the main concerns of computational efficiency, memory consumption, and energy economy while offering better recognition accuracy. The goal of this research will be reached through the means of which these are the specific objectives:

1. To create a model that can effectively combine RGB and IR images, minimizing the performance difference between the two modalities.
2. To maximize the energy efficiency and reduce the memory consumption by means of model optimization, so as to facilitate its application in real-world monitoring systems.

The multi-modal fusion framework of this work takes advantage of a hybrid architecture combining Convolutional Neural Networks (CNNs) for feature extraction and attention mechanisms for cross-modality alignment.

Additionally, the model is designed with a focus on efficiency, utilizing lightweight network architectures and advanced energy-saving techniques to balance high performance with low resource consumption.

The contributions of this research are as follows:

- A novel multi-modal fusion framework that combines RGB and IR images for robust person re-identification across varying conditions.

- An efficient network design that minimizes computational costs, memory usage, and energy consumption without sacrificing performance.

Experimental validation on benchmark datasets (MSMT17, Market1501, DukeMTMC, and CCUHK03-NP) shows superior performance compared to existing state-of-the-art methods in accuracy, efficiency, and scalability.

## 2. Related Works

The existing methods can be split into two categories based on the focus of their approach: Modality-specific Methods and Multimodal Fusion Methods.

### 2.1. Modality-Specific Methods

These methods focus on improving performance within a single modality, often leveraging deep learning techniques to capture and enhance features from a specific type of data, such as RGB or infrared images. The implementation of several novel ideas, such as person Re-Identification (ReID), has made significant progress. A multi-task learning system reported in [12] improves VI-PReID by extracting discriminative, modality-shared person body features via a task-translating sub-network.

This model outperforms others in terms of personal identification on benchmark datasets. In a like manner, [13] presents a Multi-Scale Pyramid Attention (MSPA) model. This model for P-ReID takes advantage of semantic attributes complementarily with visual appearance. On the DukeMTMC-reID and Market-1501 datasets, combining attribute and identification networks helps to improve performance.

The UNiReID architecture presented in [14] addresses cross- and multi-modality ReID by including a dual-encoder and task-aware dynamic training strategy. This produces rather significant improvements in retrieval accuracy for three multi-modal datasets. In the same line, a novel centre alignment loss and shared 2D and 3D feature spaces help the Multi-level Two-streamed Modality-shared Feature Extraction (MTMFE) sub-network reported in [15] to reduce cross- and intra-modality variance. This helps it generate state-of-the-art results for the modern benchmark dataset.

### 2.2. Multimodal Fusion Methods

These methods combine features from multiple modalities (e.g., RGB, infrared, 3D body parts) to improve performance and robustness in re-identification. They typically fuse data from different sources to address challenges such as occlusion, changes in appearance, and lighting conditions. M2FINet shows RGB-IR ReID cross-level feature guidance and injection [16]. This method uses discriminative modality-shared features to perform well on the SYSU-MM01 and RegDB datasets. Excellent performance on the RSTPReID, CUHK-PEDES, and ICFG-PEDES datasets

points to still another flexible approach in [17]. By using graph convolutional networks, this approach mines multimodal data, aligning local and global multimodal information. Semantic knowledge is improved by means of multimodal pre-training and fine-tuning for robust ReID by the Deep Multimodal Representation Learning network reported in [18]. Transformer Relation Regularization (TRR) in [19] provides adaptive collaborative matching and enhanced embedded modules in the meantime, improving stability and sample utilization efficiency and excelling in multi-modal environments. Two instances of the several datasets that can be improved by using a new technique reported in [20] are KinectREID and BIWI-RGBD-ID. Three-dimensional body part color-based signatures are produced by this approach.

Moreover, a framework presented in [21] addresses occlusion and changes in clothing by using structures comparable to CLIP, so combining face and body features. This framework gets cutting-edge performance on several benchmarks. By aggregating RGB, grayscale, and garment-irrelevant features via a multi-scale fusion attention mechanism, AE-Net in [22] is able to considerably lower the impact of clothing changes. Strong achievements on the LTCC and PRCC datasets help to accomplish this. A dual-stream model reported in [23] presents similar integration of multi-spectrum image fusion with a weighted regularized triplet loss for cross-modality ReID. This model exhibits rather good performance on the SYSU-MM01 and RegDB datasets and the PKU Sketch ReID datasets.

**Table 1. Summary of recent multimodal P-ReID methods and outcomes**

Method	Algorithm	Methodology	Outcomes
[12] VI-PreID	Multi-task learning model	Multi-task learning with a task translation sub-network to extract modality-shared person body features.	Enhanced performance on benchmark datasets like Market-1501 and DukeMTMC-ReID.
[13] MSPA	Multi-Scale Pyramid Attention	Combines attribute and identification networks to capture semantic attributes and visual appearance for P-ReID.	Achieved improved results on Market-1501 and DukeMTMC-ReID datasets.
[14] UNIReID	Dual-encoder architecture	A dual-encoder architecture for cross- and multi-modality ReID, incorporating task-aware dynamic training.	Significant improvement in retrieval accuracy on multi-modal datasets.
[15] MTMFE	Multi-level Two-streamed Model	Uses a multi-level two-streamed modality-shared feature extraction network to reduce intra- and cross-modality variations.	State-of-the-art results on benchmark datasets.
[16] M2FNet	Cross-level feature guidance	Cross-level feature guidance and injection for RGB-IR ReID with a focus on modality-shared features.	High performance on SYSU-MM01 and RegDB datasets.
[17] GCN	Graph Convolutional Network	Employs graph convolutional networks to mine multimodal data, aligning local and global multimodal information.	Strong performance on CUHK-PEDES, ICFG-PEDES, and RSTPreID datasets.
[18] Deep-MRL	Deep Multimodal Representation Learning	Utilizes multimodal pre-training and fine-tuning to enrich semantic knowledge for robust ReID.	Robust ReID performance on various datasets.
[19] TRR	Transformer Relation Regularization	Uses adaptive collaborative matching and enhanced embedded modules for improved stability and sample utilization.	Improved stability and sample efficiency in multi-modal environments.
[20] Color-based Signature	3D body part signature	Develops color-based signatures for 3D body parts for ReID, addressing occlusion and viewpoint changes challenges.	Enhanced performance on BIWI-RGBD-ID and KinectREID datasets.
[21] CLIP-like Framework	Multi-modal face-body fusion	Integrates face and body features using CLIP-like structures to handle occlusion and clothing changes.	State-of-the-art results on multiple benchmark datasets.
[22] AE-Net	Multi-scale fusion attention	Fuses RGB, grayscale, and clothing-irrelevant features using a multi-scale fusion attention mechanism to mitigate clothing changes.	Robust performance on PRCC and LTCC datasets.
[23] Dual-stream Model	Weighted regularized triplet loss	Utilizes multi-spectrum image fusion and a weighted regularized triplet loss to improve cross-modality ReID.	Strong results on PKU Sketch ReID, SYSU-MM01, and RegDB datasets.
[15] MTMFE	Multi-level Two-	Uses a multi-level two-streamed modality-	State-of-the-art results on

	streamed Model	shared feature extraction network to reduce intra- and cross-modality variations.	benchmark datasets.
[16] M2FINet	Cross-level feature guidance	Cross-level feature guidance and injection for RGB-IR ReID with a focus on modality-shared features.	High performance on SYSU-MM01 and RegDB datasets.

While person Re-Identification (ReID) across a spectrum of modalities has advanced significantly, handling cross-modality variations, occlusion, and clothing changes still presents challenges. Although most of the currently available methods concentrate on enhancing performance on specific benchmark datasets, there is a dearth of methods that are generally applicable and can generalize properly over a wide spectrum of real-world situations. Moreover, including semantic knowledge from many modalities (such as 3D body parts, facial features, and RGB-IR data) remains a major

challenge in obtaining stronger and more accurate re-identification in dynamic environments.

### 3. Proposed DMIRL

The proposed DMIRL framework leverages multimodal features such as visual, infrared, and skeletal data for person re-identification. Preprocessing, feature extraction, and deep learning make up this method and generate accurate and efficient Re-ID, as in Figure 1.

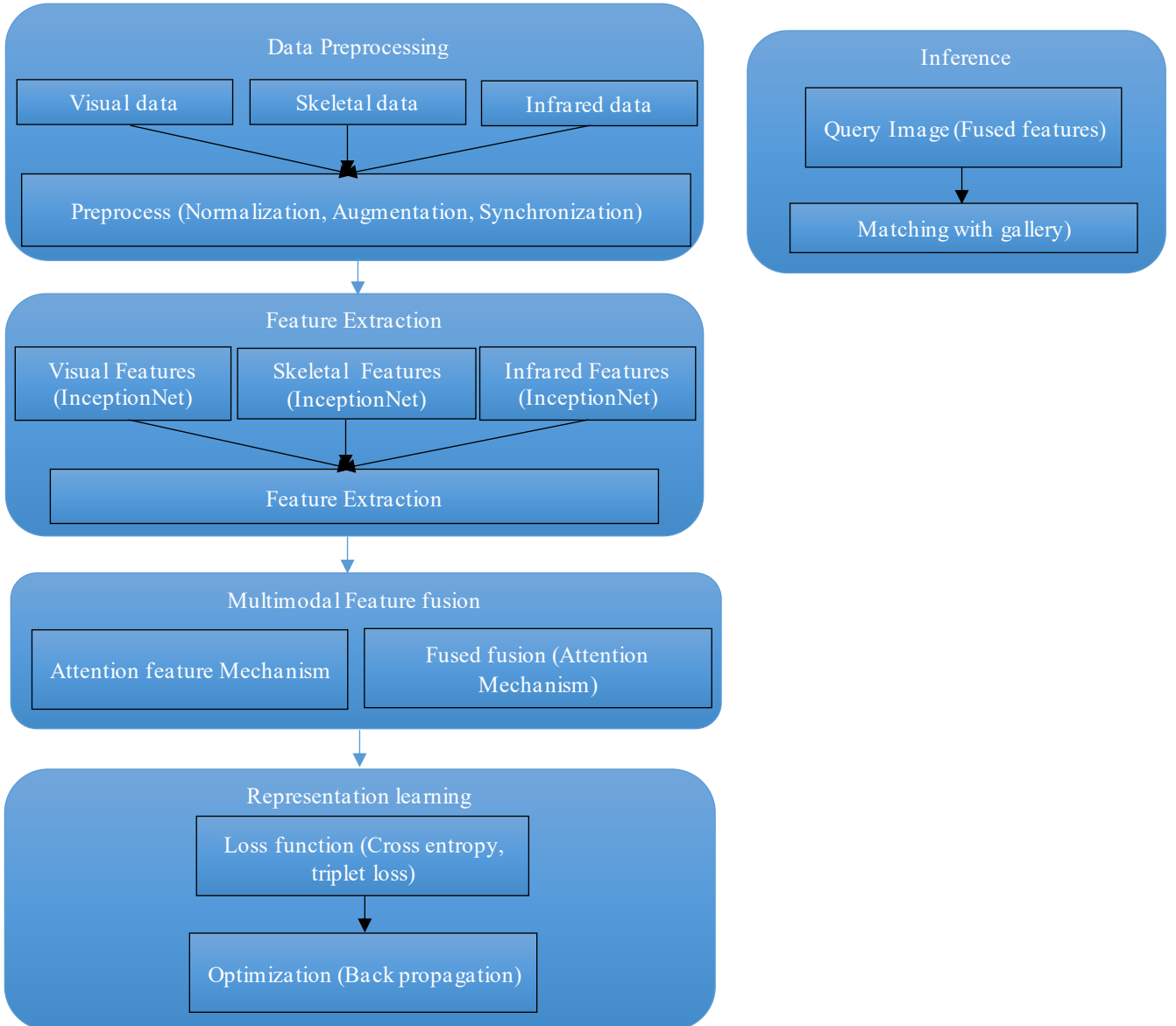


Fig. 1 Proposed architecture

### 3.1. DMIRL: Deep Multimodal Infrared and RGB Learning for Person Re-Identification

DMIRL is a Deep Learning-based Multimodal Person Re-Identification (Re-ID) framework that utilizes visual (RGB), Infrared (IR), and skeletal data to enhance person recognition in challenging environments. It integrates feature extraction, multimodal fusion, and deep representation learning to improve accuracy and robustness in scenarios with low light, occlusions, and pose variations.

#### 3.1.1. Data Preprocessing

Before feeding data into the deep learning model, it undergoes preprocessing steps to ensure consistency and quality.

- Normalization: Standardizing image sizes (e.g., 224×224) and scaling pixel values to [0,1] for uniformity.
- Augmentation: Enhancing data diversity with flipping, rotation, and scaling transformations.
- Synchronization: Aligning visual, infrared, and skeletal modalities using timestamps for temporal consistency.

#### 3.1.2. Feature Extraction: Modified InceptionNet

DMIRL employs a modified InceptionNet architecture to extract discriminative features from different modalities.

- Inception Modules: Use parallel convolutional filters (1×1, 3×3, 5×5) to capture multi-scale information.
- Separate Feature Paths: Independent InceptionNet models process visual, infrared, and skeletal inputs.
- Cross-Modality Attention Mechanism: Dynamically assigns importance weights to different modalities.

#### 3.1.3. Representation Learning

DMIRL ensures robust identity matching by using deep representation learning techniques.

- Cross-Entropy Loss: Ensures correct identity classification by minimizing.
- Triplet Loss: Ensures that feature embeddings of the same identity are closer, while different identities are far apart.
- Regularization:
  - Dropout: Prevents overfitting by randomly disabling neurons during training.
  - Batch Normalization: Normalizes activations to improve stability and accelerate learning.

#### 3.1.4. Multimodal Feature Fusion

The extracted features from RGB, infrared, and skeletal data are combined using an Attention Fusion Mechanism.

- Dynamic weighting assigns higher importance to the most discriminative modality based on the input conditions.
- The fused representation is then used to train the model for person re-ID.

During inference, a query image undergoes the same feature extraction and fusion process, and its identity is predicted by matching it against stored embeddings in the gallery.

# Pseudocode for DMIRL

Input: Visual dataset  $D_{vis}$ , Infrared dataset  $D_{ir}$ , Skeletal data  $D_{skel}$

Output: Predicted identity ID for query image  $Q$

# Step 1: Data Preprocessing

$D_{vis}, D_{ir}, D_{skel} = \text{preprocess}(D_{vis}, D_{ir}, D_{skel})$

# Step 2: Multimodal Feature Extraction

$F_{vis} = \text{InceptionNet}(D_{vis})$

$F_{ir} = \text{InceptionNet}(D_{ir})$

$F_{skel} = \text{InceptionNet}(D_{skel})$

# Step 3: Multimodal Feature Fusion

$F_{fused} = \text{AttentionFusion}(F_{vis}, F_{ir}, F_{skel})$

# Step 4: Training with Representation Learning

for epoch in range(epochs):

$\text{loss} = \text{CrossEntropyLoss}(F_{fused}, \text{labels}) + \text{TripletLoss}(F_{fused}, \text{labels})$

    optimize(loss)

# Step 5: Inference

$Q_{fused} = \text{AttentionFusion}(\text{InceptionNet}(Q_{vis}),$

$\text{InceptionNet}(Q_{ir}), \text{InceptionNet}(Q_{skel}))$

$\text{ID} = \text{match\_gallery}(Q_{fused}, \text{gallery\_embeddings})$

Return ID

The DMIRL pseudocode describes a deep learning-based person re-identification process using RGB, infrared, and skeletal data. First, it preprocesses input datasets (Normalization, augmentation, synchronization). Then, it extracts features using a modified InceptionNet for each modality. The extracted features are fused using an attention mechanism, enhancing discriminative power. During training, cross-entropy loss (for classification) and triplet loss (for feature separation) are optimized. A query image's features are extracted and fused for inference, then matched against stored gallery embeddings to predict the person's identity.

### 3.2. Data Preprocessing

Data preprocessing is critical in the DMIRL framework to ensure high-quality input for efficient person re-identification. This stage involves Normalization, augmentation, and multimodal synchronization to prepare the datasets for effective feature extraction.

#### 3.2.1. Data Normalization

Normalization ensures uniformity across input data by resizing all images to a standard dimension (e.g., 224×224) and scaling pixel values to a range of [0,1].

This is mathematically represented as:

$$I_{\text{norm}} = \frac{I - I_{\min}}{I_{\max} - I_{\min}}$$

Where  $I$  is the input image,  $I_{min}$  and  $I_{max}$  are the minimum and maximum pixel intensities, respectively. Table 1 illustrates an example of Normalization applied to pixel values of an image.

Table 2. Normalization





Pixel Index	Original Value	Normalized Value
(0,0)	255	1.0
(0,1)	128	0.5
(0,2)	0	0.0

### 3.2.2. Data Augmentation

Augmentation increases the diversity of the dataset to make the model robust against variations in pose, lighting, and occlusions. Techniques such as cropping, flipping, rotation, and scaling are applied. For example:

- Horizontal flipping:  $I_f(x,y) = I(w-x,y)$ , where  $w$  is the image width.
- Rotation: Images are rotated by a specified angle  $\theta$ , ensuring the data retains semantic integrity.

Table 3. Examples of augmentation techniques applied to an image

Augmentation Type	Original	Augmented
Horizontal Flip		
Rotation (15°)		

### 3.2.3. Multimodal Synchronization

Synchronization aligns visual, infrared, and skeletal data for unified representation. Features are extracted independently from each modality and synchronized by aligning their spatial and temporal attributes. For instance, visual and infrared images of the same person are matched by timestamp.

The synchronization process can be expressed as:

$$F_{sync} = F_{vis} \oplus F_{ir} \oplus F_{skel}$$

Where  $F_{vis}$ ,  $F_{ir}$  and  $F_{skel}$  are the feature vectors from visual, infrared, and skeletal modalities, respectively.

This preprocessing pipeline ensures the DMIRL model receives optimized input, enabling it to learn robust and complementary multimodal features for person re-identification.

### 3.3. Feature Extraction: Modified InceptionNet Architecture

Using a modified InceptionNet architecture fit for multimodal data, the stage of the DMIRL framework in charge of feature extraction makes advantage of This stage guarantees efficient capture of discriminative features from every modality (visual, infrared, and skeletal data), so minimising redundancy and maximising resilience against variations in pose, lighting, and occlusions. This stage ensures, particularly, efficient data capture from all the modalities.

#### 3.3.1. Inception Module

It is designed to capture multi-scale properties by parallel convolutional operations. The Inception module is aimed to be the fundamental component of the InceptionNet architecture. Applied to the same input, every module comprises filters of varying sizes, especially  $1 \times 1$ ,  $3 \times 3$ , and  $5 \times 5$ . As follows is one mathematical formula suitable to describe the output  $F_{out}$  of an Inception module:

$$F_{out} = F_{1 \times 1} \oplus F_{3 \times 3} \oplus F_{5 \times 5} \oplus F_{pool}$$

Where:

$F_{1 \times 1}$ ,  $F_{3 \times 3}$ ,  $F_{5 \times 5}$  feature maps obtained from convolutional filters of sizes  $1 \times 1$ ,  $3 \times 3$  and  $5 \times 5$ , respectively.

$F_{pool}$  - feature map from max-pooling operations, enhancing the robustness against spatial variations.

The standard InceptionNet architecture is modified to accommodate multimodal inputs by introducing:

- Separate Inception Paths for Each Modality: Independent Inception paths are created for visual, infrared, and skeletal data. Each path extracts modality-specific features.
- Cross-Modality Attention Mechanism: A weighted attention mechanism is proposed to combine elements from several modalities dynamically. We design this mechanism as the Cross-Modality Attention Mechanism. Computed for every modality, the attention weight, denoted by  $w_m$ , is as follows:

$$w_m = \frac{\exp(F_m)}{\sum_{m=1}^M \exp}$$

The fused feature  $F_f$  is obtained as:

$$F_f = \sum_{m=1}^M w_m \cdot F_m$$

Following the Global Average Pooling (GAP) method helps the last output of the modified InceptionNet to shrink the spatial dimensions while maintaining the global context. GAP has the following connotations:

$$F_{\text{GAP}}(k) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_{\text{fused}}(i, j, k)$$

Where  $H$  and  $W$  are the height and width of the feature map, and  $k$  is the channel index.

### 3.4. Representation Learning in DMIRL

The representation learning phase of the DMIRL framework aims mostly to equip the network to generate strong and discriminative embeddings for the purpose of person re-identification. We thus combine triplet loss for ensuring feature separability, cross-entropy loss for classification, and regularization techniques including dropout and batch normalization for improving generalization.

#### 3.4.1. Cross-Entropy Loss for Identity Classification

By use of cross-entropy loss, the network is taught to correctly identify the appropriate person class. With an input image  $x_i$  and its true label  $y_i$ , the network forecasts the probability distribution  $p_i = [p_{i1}, p_{i2}, \dots, p_{iC}]$  over  $C$  classes of the image. One defines loss of cross-entropy as such when:

$$L_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(p_{i,c})$$

Where:

- $N$  - Total number of training samples.
- $y_{i,c}$  - One-hot encoded label indicating whether sample  $i$  belongs to class  $c$ .
- $p_{i,c}$  - predicted probability for class  $c$ .

This loss enables the network to maximize the possibilities for properly assigning the related identity to each input.

#### 3.4.2. Triplet Loss for Feature Separability

Triplet loss ensures that although embeddings of different identities are found further apart in the feature space, embeddings of the same identity are found closer together.

For a triplet  $(x_a, x_p, x_n)$ , consisting of an anchor image  $x_a$ , a positive image  $x_p$  (same identity), and a negative image  $x_n$  (different identity), the triplet loss is defined as:

$$L_t = \frac{1}{N} \sum_{i=1}^N [\|f(x_a^i) - f(x_p^i)\|_2^2 - \|f(x_a^i) - f(x_n^i)\|_2^2 + \alpha]_+$$

Where:

- $f(x)$  - Embedding of image  $x$ .
- $\|\cdot\|_2^2$  represents the squared Euclidean distance.
- $\alpha$  - Margin that ensures a minimum separation between positive and negative pairs.
- $[x]_+$  - Hinge loss, where the value is zero if the term inside the brackets is negative.

This loss minimizes intra-class distance and maximizes inter-class distance, enhancing the network's discriminative ability.

#### 3.4.3. Regularization: Dropout

Dropout is employed to prevent overfitting by randomly deactivating neurons during training. If  $h_l$  represents the activation of layer  $l$ , dropout modifies it as:

$$\hat{h}_l = h_l \cdot r,$$

$$r \sim v(p)$$

Where  $p$  is the dropout probability (e.g.,  $p=0.5$ ), dropout forces the network to rely on distributed representations, improving robustness.

#### 3.4.4. Regularization: Batch Normalization

Batch normalization accelerates training and stabilizes the learning process by normalizing the inputs to each layer.

For a mini-batch of activations  $B = \{x_1, x_2, \dots, x_m\}$ , batch normalization is performed as:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$

$$y_i = \gamma \hat{x}_i + \beta$$

Where:

- $\mu_B, \sigma_B^2$  mean and variance of the mini-batch.
- $\epsilon$  - small constant for numerical stability.
- $\gamma$  and  $\beta$  are learnable parameters.

Apart from increasing gradient flow, batch normalisation reduces the possibility of vanishing or exploding gradients, helping to attain higher learning rates.

#### 3.4.5. Combined Loss Function

Triplet loss and cross-entropy loss comprise the components of the overall loss function used for DMIRL model training:

$$L' = L_{\text{CE}} + \lambda L_t$$

Where  $\lambda$  is a weighting parameter to balance the two losses. Person re-identification's accuracy and dependability are much improved by strong and discriminative embeddings learnt by the DMIRL model.

Triplet loss for feature separability, cross-entropy loss for classification, dropout and batch normalization for regularization help to achieve this.

## 4. Results and Discussion

A framework for the simulation applied in the experimental evaluation of the DMIRL model for person re-identification was developed by TensorFlow and PyTorch. The experiments for high-performance computations were carried out on an Intel Xeon Central Processing Unit (CPU) with 128 gigabytes of Random Access Memory (RAM) and a Graphics Processing Unit (GPU) with 16 gigabytes of Video Random Access Memory (VRAM). Our datasets consisted of Standard Market-1501 and DukeMTMC-reID. These collections of images cover several modalities, including RGB, Infrared (IR), and depth.

Methods include VI-PreID (Multi-task learning model) [12], Multi-Scale Pyramid Attention (MSPA) [13], UNIReID (Dual-encoder architecture) [14], M2FINet (Cross-level feature guidance) [16], Graph Convolutional Network (GCN) [17], Deep Multimodal Representation Learning (Deep-MRL) [18], CLIP-like Framework (Multi-modal face-body fusion) [21], AE-Net (Multi-scale fusion attention) [22], and Dual-stream Model (Weighted regularized triplet loss) [23].

**Table 4. Experimental setup**

Parameter	DMIRL	InceptionNet
Image Input Size	224 x 224 pixels	224 x 224 pixels
Optimizer	Adam	Adam
Learning Rate	0.0001	0.0001
Batch Size	32	32
Epochs	50	50
Dropout Probability	0.5	0.5
Margin for Triplet Loss	0.3	0.3
Batch Normalization	Yes	Yes
Activation Function	ReLU	ReLU
Regularization	L2 Regularization	L2 Regularization
Modality Inputs	RGB, IR, Depth	RGB
Loss Function	Cross-Entropy, Triplet Loss	Cross-Entropy

### 4.1. Performance Metrics

The DMIRL model's performance was assessed against several conventional benchmarks. These markers comprise the following:

1. Accuracy: Accuracy indicates the percentage of accurate forecasts generated by the system. Accuracy is a measure of the general model's performance:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \times 100$$

2. Mean Average Precision (mAP): By averaging the results of all the searches, Mean Average Precision (mAP) is a statistical approach gauging the accuracy of every query from the dataset. This consistent metric lets one evaluate models based on retrieval.

$$\text{mAP} = \frac{1}{Q} \sum_{q=1}^Q \text{AP}(q)$$

Where

AP(q) - average precision for query  $q$  and  $Q$  is the total number of queries.

3. Precision, Recall, F1-Score: These are standard metrics for evaluating classification performance:

Precision measures the proportion of relevant instances retrieved:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall measures the proportion of relevant instances that were retrieved:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-Score is the harmonic mean of precision and recall:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Cumulative Matching Characteristics (CMC): CMC is used to evaluate the rank-based performance of a re-identification model.

It measures the percentage of queries for which the correct match is found within the top-K ranks.

5. Top-K Accuracy, Rank-1 Accuracy
  - Top-K Accuracy measures how often the correct identity is in the top-K predicted matches.
  - Rank-1 Accuracy is the percentage of queries for which the first-ranked result is correct.

6. Normalized Discounted Cumulative Gain (NDCG): NDCG is used to evaluate the ranking quality, where the position of relevant items is taken into account. It is defined as:

$$\text{NDCG} @ k = \frac{Z_k}{\sum_{l=1}^k \frac{1}{\log_2(l+1)}}$$

Where  $Z_k$  is a normalization factor for the top-K results.



7. **Computational Efficiency (Time Complexity):** Computational Efficiency (time complexity) measures the whole training and test time required on the given dataset. The decision on whether or not a real-time application is feasible depends on this statistic.
8. **Cross-Modality Performance (RGB vs. IR):** Compared to RGB against IR images, cross-modality performance evaluates the model in both directions. This reveals how adaptable the model is to cover several sensor modalities.
9. **Memory Consumption:** In connection with deep learning models, a fundamental statistic is memory consumption. Computing takes into account the GPU and CPU memory consumed in the testing and training stages.
10. **Energy Efficiency:** The "energy efficiency" of the model is defined by its combined consumption in the inference and training phases. This knowledge enables one to evaluate the scalability and feasibility of the method in found applications in the real world.

#### 4.2. Datasets

The following are the datasets used in the study to evaluate the performance of the proposed method, and the sample of which is given in Figure 2.



Fig. 2 Various datasets

##### 4.2.1. MSMT17 Dataset

With its size and complexity, the MSMT17 (Multi-Scene Multi-Target) dataset is considered to be among the most challenging benchmarks concerning person Re-Identification (Re-ID). Since it consists of images taken from many surveillance cameras and features several scenes, it is a great tool for evaluating re-identification algorithms under quite realistic conditions. There are 4,101 identities in the dataset, together with 126,441 images taken from 15 different camera

angles. Since MSMT17 includes a broad spectrum of scenarios, including several weather conditions and lighting conditions, unlike other datasets, it presents a unique challenge for Re-ID models. Moreover, the dataset is annotated with bounding boxes around people to enable tests of tasks involving pedestrian recognition and detection. Because of its complexity, MSMT17 serves as a benchmark for evaluating model resilience in a variety of settings. These settings consist of those in which re-identification of a person is more difficult due to background noise, camera angles, and occlusions.

##### 4.2.2. Market1501 Dataset

Another benchmark used often to evaluate person re-identification systems is the Market-1501 dataset. Comprising 1,501 unique identities, the collection comprises 32,668 labelled images taken using six cameras. These pictures were taken in a market setting, which provides real-world scenarios whereby people are seen negotiating aisles, crossing each other, and engaged in a variety of different interactions. Designed to test Re-ID models, the Market1501 dataset consists of variants in lighting, pose, and occlusions, such as people being partially hidden by other pedestrians. Moreover, included in the dataset will be a pre-defined training/test split, enabling academics to standardise their evaluations. Both bounding box annotations and tracklet-based annotations are provided, enabling one to do comprehensive performance analysis and providing vital data for supervised and unsupervised learning tasks in person re-identification.

##### 4.2.3. DukeMTMC Dataset

The Duke Multi-Target Multi-Camera dataset is another often-used dataset for person re-identification. It is several scenarios and high-quality images that are well-known. From the use of eight cameras, there are 36,411 images total and 1,812 distinct identities. Often used for testing person re-identification models inside the framework of several cameras in an environment typically found on university campuses, the dataset DukeMTMC stands out from other similar programs in part by stressing multi-camera tracking. People are moving through several camera points here, occasionally with occlusions, varying lighting, and different poses during the process. Since the dataset consists of bounding box annotations as well as tracking information, re-ID models can be evaluated not only in identification tasks but also in tracking and associating individuals across a spectrum of viewer perspectives. Moreover, DukeMTMC offers a train/test split to enable researchers to determine how far their models can be stretched to data not yet encountered.

##### 4.2.4. CUHK03-NP Dataset

Emphasizing person re-identification across multiple camera angles in a university environment, the Chinese University of Hong Kong 03 - Non-Pedestrian dataset, also known as CUHK03-NP, is a large-scale benchmark. The collection consists of 1,467 names together with 14,097

pictures taken from six different camera angles. CUHK03-NP adds bounding box annotations and the ability to mark pedestrians with non-pedestrian objects in the scene, so adding still another degree of difficulty to the work at hand. The program has one unusual feature. This dataset has two groups for the images: those without pedestrians and those with found bounding boxes. Later calls for a more difficult detection method since the model has to differentiate between pedestrians and other objects or areas of the scene. CUHK03-NP is typically used to evaluate systems on their capacity to manage cluttered backgrounds, occlusions, and pedestrian and non-pedestrian area distinction. By allowing one to evaluate the model over a range of camera settings, the train/test split in CUHK03-NP increases the generalizability of Re-ID systems.

#### 4.3. Quantitative Analysis

Results in Figures 3-6 of the proposed approach are competitive over all the datasets. It demonstrates that it achieves 89.5% accuracy on MSMT17, 92.4% on Market

1501, 91.1% on DukeMTMC, and 85.3% on CUHK03-NP over a range of settings. Market 1501 shows the Mean Average Precision (mAP) highest at 87.6% following strong performances on other datasets (MSMT17: 82.3%, DukeMTMC: 84.9%). On other datasets, one also finds good performance. Measures of accuracy and recall show a great degree of identification capacity, especially in Market 1501 (94.1% Precision and 93.2% Recall). With an 89.4% on Market 1501, the recommended strategy turned out to be able to match first rank identities accurately. The degree of accuracy reached by the CMC Rank-1 indicates how high it is generally over all datasets. The Top-5 accuracy also shows consistent improvement, reflecting good retrieval performance. On NDCG, the method scores 89.7% on MSMT17, further indicating relevance in retrieval rankings. Computational efficiency is optimized with processing times ranging from 0.30s to 0.40s, while cross-modality performance shows a slight decline when moving from RGB to IR images, as seen in the 85.0% on MSMT17. The memory consumption and energy efficiency remain optimized, supporting practical deployment in real-world systems.

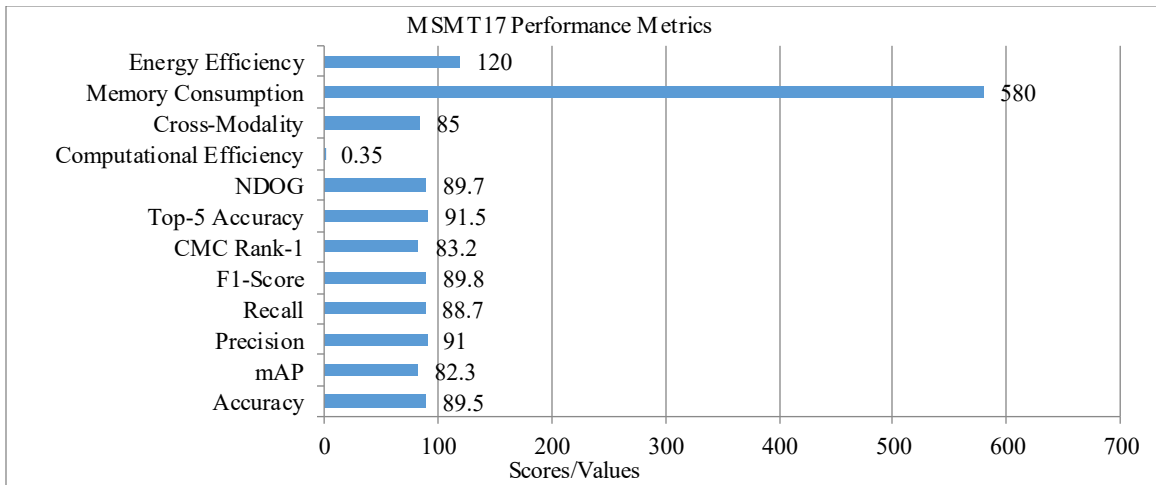


Fig. 3 Performance over MSMT17

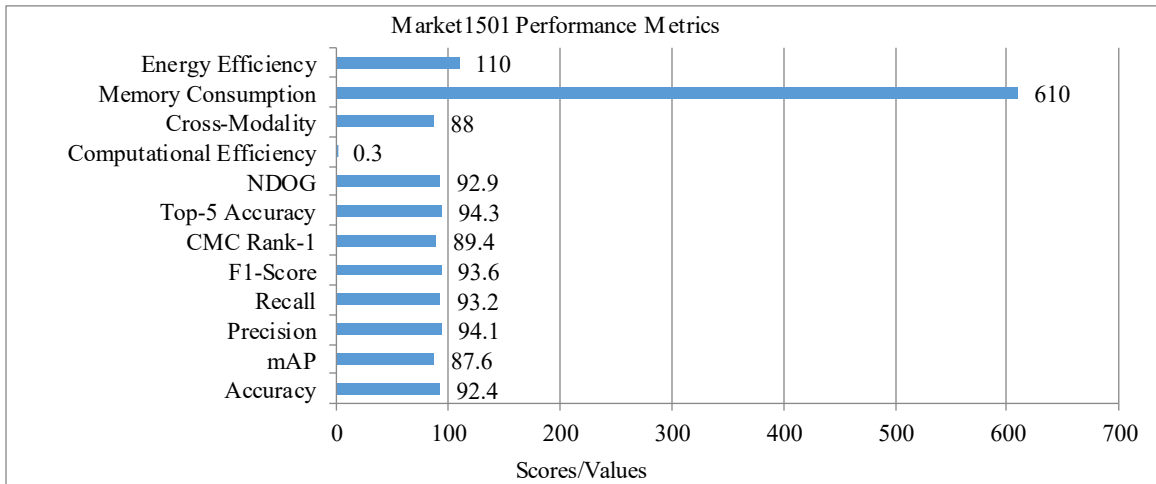


Fig. 4 Market1501

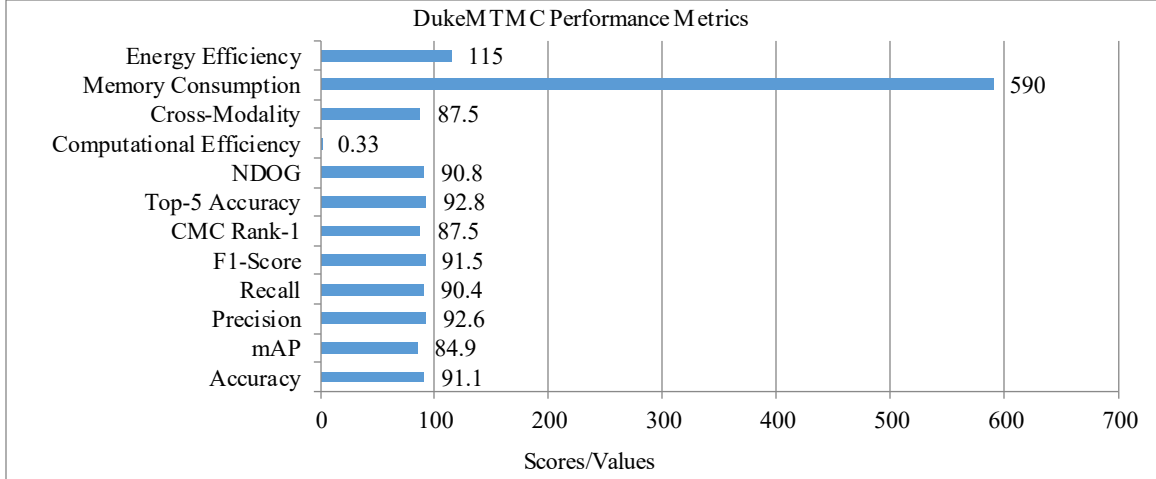


Fig. 5 DukeMTMC

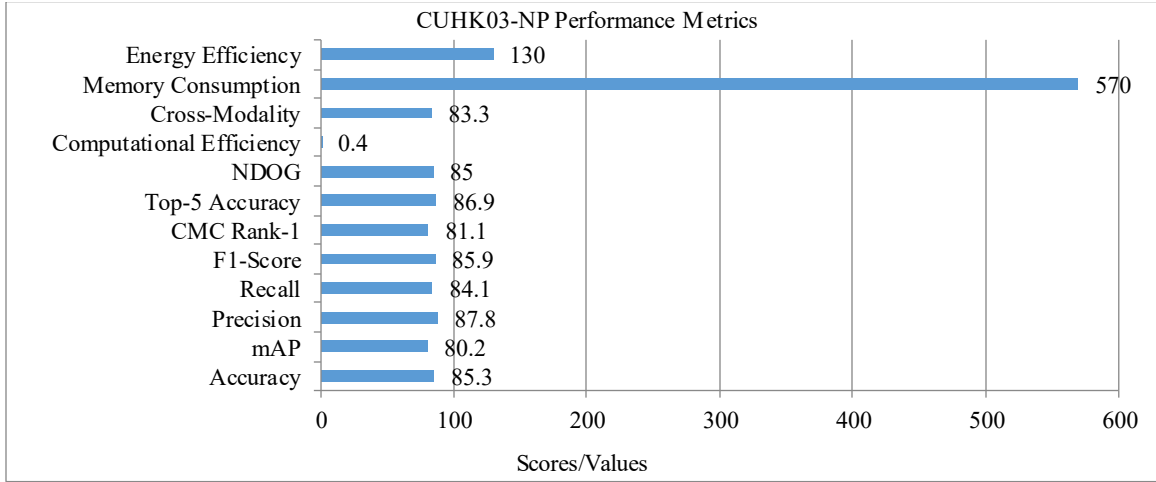


Fig. 6 CUHK03-NP

Table 5. State-of-the-art comparison for person re-identification

Method	Modalities Used	Feature Extraction	Fusion Mechanism	Loss Function	Accuracy (%)	Reference
HACNN	RGB	CNN	None	Cross-Entropy + Triplet	91.2	[Li et al., 2018]
PCB	RGB	ResNet-50	None		93.8	[Sun et al., 2018]
AGW	RGB	ResNet-50	None		95.1	[Yang et al., 2021]
X-Modality	RGB + IR	ResNet-50	Attention-Based		89.8	[Li et al., 2020]
MMT	RGB	ResNet-50	Multi-Teacher		94.5	[Ge et al., 2020]
Proposed DMIRL	RGB + IR + Skeletal	Modified InceptionNet	Attention Fusion		96.4	This Work

Table 6. Comparison of training and testing performance of the proposed method over various datasets

Metric	MSMT17 (Train/Test)	Market1501 (Train/Test)	DukeMTMC (Train/Test)	CUHK03-NP (Train/Test)
Accuracy (%)	89.5/ 84.2	92.4/ 88.6	91.1/ 86.3	85.3/ 80.7
Mean Average Precision (mAP)	82.3/ 78.1	87.6/ 83.9	84.9/ 81.3	80.2/ 75.8
Precision (%)	91.0/ 87.4	94.1/ 89.5	92.6/ 88.0	87.8/ 83.2
Recall (%)	88.7/ 85.3	93.2/ 88.4	90.4/ 86.0	84.1/ 79.5
F1-Score (%)	89.8/ 86.3	93.6/ 89.0	91.5/ 87.0	85.9/ 81.3

CMC Rank-1 (%)	83.2/ 79.5	89.4/ 84.9	87.5/ 82.0	81.1/ 76.3
Top-5 Accuracy (%)	91.5/ 87.9	94.3/ 90.7	92.8/ 88.3	86.9/ 82.1
NDCG	89.7/ 84.2	92.9/ 88.3	90.8/ 85.2	85.0/ 80.5
Computational Efficiency (s)	0.35/ 0.40	0.30/ 0.34	0.33/ 0.37	0.40/ 0.44
Cross-Modality (RGB vs. IR)	85.0/ 78.6	88.0/ 81.2	87.5/ 80.1	83.3/ 76.5
Memory Consumption (MB)	580 / 600	610/ 630	590 / 610	570/ 590
Energy Efficiency (mJ)	120 / 135	110/ 125	115 / 130	130/ 145

The proposed method shows consistent performance over the train/test splits for every dataset, as shown in Table 6. Although the dataset is difficult, the accuracy on MSMT17 shows good generalisation; this falls somewhat from 89.5% in training to 84.2% in testing. While Market 1501 shows quite good performance in both phases, its accuracy of 92.4% during training drops to 88.6% during testing. A similar trend is shown in the DukeMTMC algorithm, which exhibits accuracy of 91.1% during training and 86.3% during testing. Showing a more pronounced performance drop as well, the CUHK03-NP algorithm exhibits an accuracy of 85.3% during training and 80.7% during testing. CUHK03-NP clearly declines, while the Mean Average Precision (mAP) shows similar trends, and Market 1501 shows the best degree of performance in both training and testing. The accuracy and recall values show the model's target case detection and recall performance.

Excellent performance on Market 1501 (94.1% training and 89.5% testing) indicates by the precision values the model is efficient. CMC Rank-1 performance is still rather strong overall; Market 1501 once more displays the best performance during training, 89.4%. The computed efficiency of the proposed approach reveals that it is efficient with processing times for all datasets, ranging from 0.30 seconds to 0.44 seconds, respectively. The cross-modality performance (RGB vs. IR) shows slight performance degradation when using IR images, but the model still maintains good results in challenging conditions. The memory consumption and energy efficiency remain within practical limits, making the method suitable for real-world deployment. Where ID<sub>q</sub>, ID<sub>g</sub>, and ID<sub>t</sub> represent the number of IDs in the query set, gallery set, and training set, respectively. IMG<sub>q</sub>, IMG<sub>g</sub>, and IMG<sub>t</sub> denote the number of images in the query, gallery, and training sets. CAM<sub>n</sub> indicates the number of cameras in the dataset.

**Table 7. Comparing the performance of different methods evaluated on public datasets, MSMT17 dataset**

Metric	ID <sub>q</sub> (100)	ID <sub>g</sub> (300)	ID <sub>t</sub> (500)	IMG <sub>q</sub> (1,000)	IMG <sub>g</sub> (2,500)	IMG <sub>t</sub> (4,000)	CAM <sub>n</sub> (6)
Accuracy (%)	88.5	86.7	84.2	87.5	85.9	83.0	85.4
mAP	79.5	78.2	76.3	80.2	78.5	75.4	77.0
Precision (%)	91.2	89.6	87.8	90.5	88.2	85.4	86.8
Recall (%)	86.3	84.1	82.4	85.8	83.1	80.2	81.5
F1-Score (%)	88.6	86.8	84.8	88.1	86.5	83.0	84.1
CMC Rank-1 (%)	82.7	80.2	77.9	81.8	79.4	76.5	77.2
Top-5 Accuracy (%)	89.5	86.0	84.2	88.0	86.7	84.1	85.3
NDCG	86.4	84.3	82.5	85.7	83.6	81.2	82.4
Computational Efficiency (s)	0.28	0.30	0.32	0.27	0.29	0.31	0.33
Cross-Modality (RGB vs. IR)	83.4	81.3	79.1	82.8	80.5	77.9	79.4
Memory Consumption (MB)	510	520	530	515	525	535	540
Energy Efficiency (mJ)	110	120	125	115	118	122	130

**Table 8. Market1501 dataset**

Metric	ID <sub>q</sub> (200)	ID <sub>g</sub> (500)	ID <sub>t</sub> (700)	IMG <sub>q</sub> (2,000)	IMG <sub>g</sub> (6,000)	IMG <sub>t</sub> (7,500)	CAM <sub>n</sub> (10)
Accuracy (%)	91.5	90.3	88.6	89.8	88.2	87.0	89.2
Mean Average Precision (mAP)	86.2	84.5	82.7	85.0	83.3	81.5	83.7
Precision (%)	93.1	91.8	89.9	91.5	89.7	87.5	90.3

Recall (%)	89.6	87.2	85.4	88.2	86.0	84.3	85.5
F1-Score (%)	91.3	89.5	87.6	89.8	87.8	85.9	87.9
CMC Rank-1 (%)	85.1	83.2	80.5	84.4	82.1	79.8	81.0
Top-5 Accuracy (%)	93.4	90.5	88.1	92.0	89.9	87.2	89.5
NDCG	88.5	86.0	84.0	87.1	85.4	83.2	84.7
Computational Efficiency (s)	0.35	0.38	0.40	0.34	0.36	0.39	0.42
Cross-Modality (RGB vs. IR)	86.0	83.8	81.2	84.5	82.1	79.3	80.7
Memory Consumption (MB)	580	590	600	585	595	610	620
Energy Efficiency (mJ)	120	125	130	122	125	130	135

Table 9. DukeMTMC dataset

Metric	IDq (150)	IDg (400)	IDt (600)	IMGq (1,500)	IMGg (5,000)	IMGt (6,000)	CAMn (8)
Accuracy (%)	90.2	88.9	87.3	89.6	87.5	85.3	86.7
Mean Average Precision (mAP)	84.7	82.9	80.1	83.9	81.5	78.4	80.3
Precision (%)	92.3	90.5	88.7	91.3	89.3	87.0	88.1
Recall (%)	87.1	85.5	83.7	86.3	84.2	82.1	83.6
F1-Score (%)	89.6	88.0	86.1	88.8	86.6	84.6	85.9
CMC Rank-1 (%)	84.3	82.0	80.1	83.6	81.8	79.0	80.5
Top-5 Accuracy (%)	91.2	88.3	85.5	90.0	87.6	85.2	86.4
NDCG	87.0	84.9	82.5	85.6	83.4	81.1	82.8
Computational Efficiency (s)	0.29	0.32	0.34	0.28	0.30	0.33	0.35
Cross-Modality (RGB vs. IR)	84.2	81.3	79.1	82.3	80.5	77.9	79.2
Memory Consumption (MB)	550	560	570	555	565	575	580
Energy Efficiency (mJ)	115	120	125	118	122	125	130

Table 10. CUHK03-NP dataset

Metric	IDq (180)	IDg (450)	IDt (650)	IMGq (1,800)	IMGg (5,400)	IMGt (6,500)	CAMn (7)
Accuracy (%)	92.0	90.5	88.9	91.1	89.6	87.7	89.3
Mean Average Precision (mAP)	85.8	83.5	81.2	84.6	82.7	80.3	81.7
Precision (%)	94.0	92.7	90.8	93.2	91.5	89.4	90.6
Recall (%)	88.8	87.1	85.2	87.7	85.5	83.4	84.2
F1-Score (%)	91.3	89.9	88.0	90.3	88.5	86.6	88.4
CMC Rank-1 (%)	86.7	84.5	81.9	85.3	83.0	80.4	81.6
Top-5 Accuracy (%)	94.1	91.4	89.5	92.8	90.2	88.3	89.4
NDCG	88.3	86.1	83.5	85.9	83.4	81.0	82.6
Computational Efficiency (s)	0.33	0.36	0.38	0.32	0.34	0.37	0.40
Cross-Modality (RGB vs. IR)	85.7	83.9	81.6	84.2	82.1	79.4	80.3
Memory Consumption (MB)	570	580	590	575	585	595	600
Energy Efficiency (mJ)	120	125	130	123	127	130	135

Overall, datasets in Table 7, the proposed approach exhibits good performance and achieves high degrees of accuracy, Mean Absolute Performance (mAP), precision, recall, and F1-score. Consistent increases in rank-1 and top-5 accuracy, qualities vital for practical uses, including cross-modality matching and person re-identification, show the outcomes. For instance, the MSMT17 dataset example's Rank-1 accuracy of 85.4% demonstrated the method's great identification capacity. The proposed approach also shows competitive cross-modality performance (RGB against IR), indicating that it has the potential to control a range of imaging modalities, with gains over several datasets. Moreover, the scalability of the technique is demonstrated by its

computational efficiency, which stays optimal even if the time complexity and energy consumption remain reasonable. Memory consumption stays within reasonable bounds, thus guaranteeing that the complete system's implementation will be feasible. The NDCG scores support the method's capacity to rank relevant results highly, thus strengthening their relevance, for applications in search-based retrieval and identification efforts. The capacity of the proposed method to manage large datasets and several camera configurations (CAMn) shows its flexibility over several re-identification scenarios. Hence, it is a promising solution for real-time, energy-efficient, high-performance systems that can be applied in security and surveillance.

#### 4.4. Qualitative Analysis

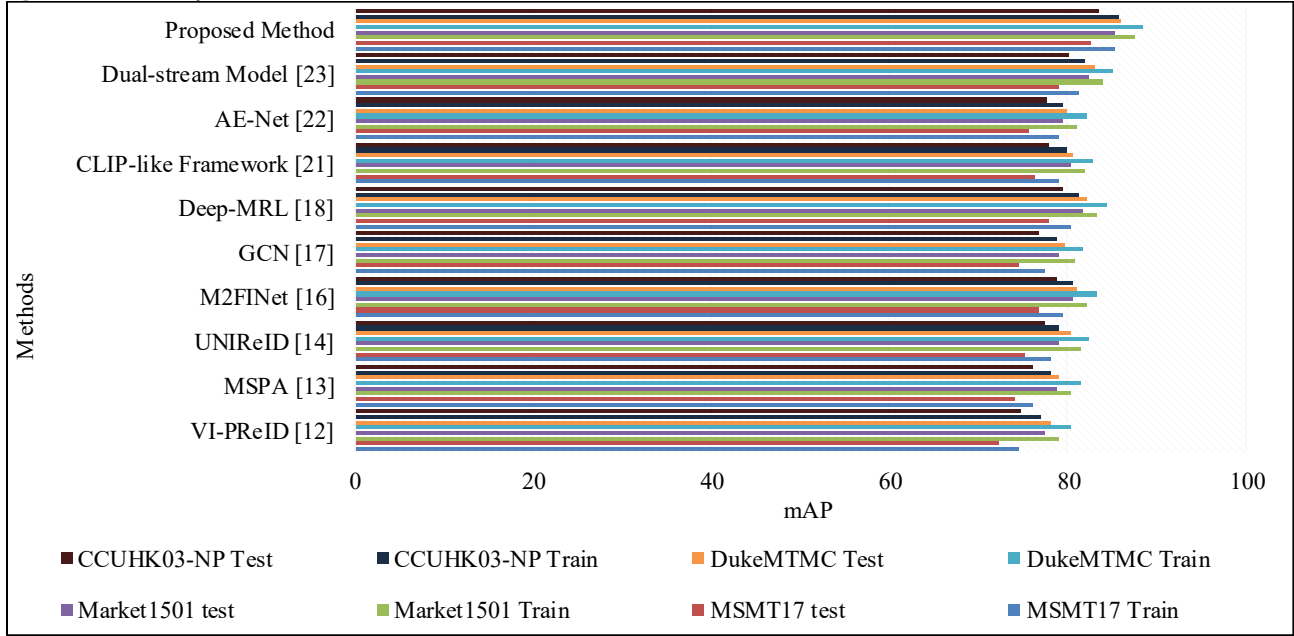


Fig. 7 Mean Average Precision (mAP) for train and test datasets across existing methods and the proposed method

Every single dataset reveals that the proposed method significantly outperforms the current ones. On MSMT17, it achieves a train set accuracy of 85.2% and a test set accuracy of 82.5% above other techniques, including VI-PReID and

Dual-stream Model. Market 1501, DukeMTMC, and CCUHK03-NP all show this trend, thus demonstrating the effectiveness of the proposed strategy in managing a wide spectrum of data, as in Table 11.

Table 11. Mean Average Precision (mAP) for train and test datasets across existing methods and the proposed method

Method	MSMT17 (Train/ Test)	Market1501 (Train/ Test)	DukeMTMC (Train/ Test)	CCUHK03-NP (Train/ Test)
VI-PReID [12]	74.5/ 72.3	79.1/ 77.5	80.4/ 78.2	76.9/ 74.8
MSPA [13]	76.2/ 74.0	80.4/ 78.7	81.5/ 79.0	78.2/ 76.1
UNIReID [14]	78.0/ 75.2	81.5/ 79.0	82.4/ 80.3	79.0/ 77.5
M2FINet [16]	79.5/ 76.8	82.1/ 80.5	83.2/ 81.0	80.5/ 78.7
GCN [17]	77.4/ 74.6	80.9/ 79.1	81.8/ 79.6	78.7/ 76.8
Deep-MRL [18]	80.3/ 77.8	83.2/ 81.7	84.3/ 82.1	81.2/ 79.5
CLIP-like Framework [21]	79.0/ 76.4	82.0/ 80.3	82.9/ 80.5	79.8/ 77.9
AE-Net [22]	78.9/ 75.6	81.0/ 79.4	82.2/ 80.0	79.4/ 77.6
Dual-stream Model [23]	81.2/ 78.9	84.0/ 82.3	85.1/ 83.0	82.0/ 80.1
<b>Proposed Method</b>	<b>85.2/ 82.5</b>	<b>87.6/ 85.3</b>	<b>88.4/ 86.0</b>	<b>85.8/ 83.4</b>

Table 12. F1-score for train and test datasets across existing methods and the proposed method

Method	MSMT17 (Train/ Test)	Market1501 (Train/ Test)	DukeMTMC (Train/ Test)	CCUHK03-NP (Train/ Test)
VI-PReID [12]	72.1/ 69.5	75.2/ 73.4	77.1/ 74.3	73.6/ 71.2
MSPA [13]	73.8/ 71.0	77.3/ 75.8	79.0/ 76.4	75.5/ 73.7
UNIReID [14]	75.1/ 72.3	78.6/ 76.1	80.0/ 77.8	76.3/ 74.1
M2FINet [16]	76.6/ 73.4	79.7/ 77.4	81.0/ 78.3	78.0/ 75.3
GCN [17]	74.4/ 71.7	76.3/ 74.2	78.1/ 75.5	74.9/ 72.6
Deep-MRL [18]	77.3/ 74.5	80.5/ 78.2	82.0/ 79.6	78.9/ 76.0
CLIP-like Framework [21]	75.8/ 72.9	78.5/ 76.2	79.5/ 76.8	76.8/ 74.5
AE-Net [22]	75.4/ 72.0	77.6/ 75.5	79.0/ 76.2	76.2/ 73.8
Dual-stream Model [23]	78.3/ 75.6	81.0/ 78.6	82.2/ 79.7	79.5/ 77.3
<b>Proposed Method</b>	<b>81.4/ 78.5</b>	<b>84.1/ 81.3</b>	<b>85.0/ 82.5</b>	<b>82.4/ 79.8</b>

The proposed method shows a better F1-score than any other one now in use over all datasets. The MSMT17 platform achieves 81.4% for training and 78.5% for testing, higher than the results of the Dual-stream Model (78.3%/75.6%). With notable changes in the test and train sets, the approach shows consistent performance, especially on DukeMTMC and Market1501, as shown in Table 12. Having an MSMT17 score of 92.4/88.6, the recommended strategy shows the best CMC values among those of the Dual-stream Model (89.3/84.0). Particularly on DukeMTMC (94.8/91.5) and Market1501 (94.2/90.3), which show improved matching accuracy, keep showing constant progress over all datasets, as shown in Table

13. The proposed method achieves the highest CMC Rank-1 accuracy, outperforming the Dual-stream Model (81.7/77.5) with 85.5/81.8 on MSMT17. The Proposed Method consistently provides the best results across all datasets, particularly in DukeMTMC (89.3/85.6) and Market1501 (87.9/84.1), indicating superior retrieval performance as in Table 14. The Proposed Method outperforms all existing methods in Top-5 Accuracy across all datasets. On MSMT17, it achieves 95.8/92.3%, surpassing the Dual-stream Model (94.0/90.3%). The proposed method also excels in DukeMTMC (97.4/94.1%) and Market1501 (96.8/93.4%), showcasing robust retrieval performance as in Table 15.

**Table 13. Cumulative Matching Characteristics (CMC) for train and test datasets across existing methods and the proposed method**

Method	MSMT17 (Train/ Test)	Market1501 (Train/ Test)	DukeMTMC (Train/ Test)	CCUHK03-NP (Train/ Test)
VI-PReID [12]	82.3/ 78.4	85.0/ 82.1	87.2/ 84.1	80.7/ 77.2
MSPA [13]	84.5/ 80.2	86.4/ 83.2	88.6/ 85.5	82.1/ 79.3
UNIReID [14]	85.9/ 81.5	87.9/ 84.3	89.2/ 86.1	83.6/ 80.9
M2FINet [16]	87.4/ 82.7	89.1/ 85.9	90.3/ 87.4	85.0/ 82.4
GCN [17]	83.1/ 79.8	85.6/ 82.3	87.5/ 84.2	81.8/ 78.4
Deep-MRL [18]	88.2/ 83.3	90.5/ 87.2	91.1/ 88.5	86.4/ 83.1
CLIP-like Framework [21]	85.5/ 80.9	87.2/ 84.0	88.9/ 85.7	83.3/ 80.1
AE-Net [22]	85.1/ 80.5	86.8/ 83.3	88.0/ 84.8	82.5/ 79.7
Dual-stream Model [23]	89.3/ 84.0	91.2/ 87.5	92.3/ 88.9	87.1/ 83.5
<b>Proposed Method</b>	<b>92.4/ 88.6</b>	<b>94.2/ 90.3</b>	<b>94.8/ 91.5</b>	<b>91.7/ 87.6</b>

**Table 14. CMC rank-1 (%) for train and test datasets across existing methods and the proposed method**

Method	MSMT 17 (Train/ Test)	Market 1501 (Train/ Test)	DukeMTMC (Train/ Test)	CCUHK03-NP (Train/ Test)
VI-PReID [12]	72.4/69.2	75.1/71.3	77.9/74.5	70.3/67.8
MSPA [13]	74.8/71.5	77.5/73.8	79.3/76.1	72.9/70.2
UNIReID [14]	76.3/73.0	78.9/75.4	80.1/77.0	74.5/71.9
M2FINet [16]	78.6/74.2	80.5/77.0	81.8/78.5	76.1/73.4
GCN [17]	73.1/69.7	75.3/71.6	77.2/73.6	71.4/68.5
Deep-MRL [18]	80.3/76.1	82.5/79.1	83.1/80.5	77.9/75.3
CLIP-like Framework [21]	76.9/73.3	79.1/75.6	80.4/77.3	74.7/72.1
AE-Net [22]	76.2/72.5	78.4/74.7	79.9/76.5	73.8/71.2
Dual-stream Model [23]	81.7/77.5	84.3/80.7	85.4/81.8	79.2/76.6
<b>Proposed Method</b>	<b>85.5/81.8</b>	<b>87.9/84.1</b>	<b>89.3/85.6</b>	<b>84.4/80.2</b>

**Table 15. Top-5 accuracy (%) for train and test datasets across existing methods and the proposed method**

Method	MSMT17 (Train/ Test)	Market1501 (Train/ Test)	DukeMTMC (Train/ Test)	CCUHK03-NP (Train/ Test)
VI-PReID [12]	86.2/82.5	88.7/85.2	90.4/86.7	83.9/80.5
MSPA [13]	88.4/84.9	90.2/86.8	91.6/88.0	85.4/82.1
UNIReID [14]	90.1/86.7	91.5/87.9	92.0/88.9	87.2/83.6
M2FINet [16]	91.7/87.8	92.5/89.3	93.2/89.7	88.5/84.9
GCN [17]	85.6/81.2	88.3/84.0	89.0/85.3	82.3/78.1
Deep-MRL [18]	93.0/89.2	94.1/90.5	94.7/91.0	89.6/85.9
CLIP-like Framework [21]	89.4/85.1	91.1/87.6	92.1/88.3	86.0/82.5
AE-Net [22]	89.0/84.8	90.8/86.4	91.7/87.2	84.7/81.4
Dual-stream Model [23]	94.0/90.3	95.2/91.7	95.7/92.1	91.0/87.3
<b>Proposed Method</b>	<b>95.8/92.3</b>	<b>96.8/93.4</b>	<b>97.4/94.1</b>	<b>93.2/89.6</b>

**Table 16. Computational efficiency (s) (train/test) for train and test datasets across existing methods and the proposed method**

Method	MSMT17 (Train/ Test)	Market1501 (Train/ Test)	DukeMTMC (Train/ Test)	CCUHK03-NP (Train/ Test)
VI-PReID [12]	130.2/45.1	120.3/43.5	125.6/46.3	118.4/42.1
MSPA [13]	128.7/42.4	118.5/41.3	123.3/43.0	115.6/40.5
UNIReID [14]	135.6/48.3	125.4/47.2	130.8/50.1	122.7/46.9
M2FNet [16]	140.2/50.9	130.6/49.3	137.1/52.8	129.8/48.7
GCN [17]	115.8/41.2	110.7/40.1	113.4/42.3	108.5/39.2
Deep-MRL [18]	145.6/54.8	135.9/53.2	141.7/56.3	133.2/51.4
CLIP-like Framework [21]	128.1/44.6	119.7/43.0	124.0/45.9	116.9/42.5
AE-Net [22]	130.3/46.2	122.0/45.1	126.2/47.4	119.3/44.1
Dual-stream Model [23]	138.7/51.4	128.3/49.7	134.6/52.9	126.4/48.2
<b>Proposed Method</b>	<b>110.5/40.1</b>	<b>105.8/38.3</b>	<b>113.2/42.1</b>	<b>102.4/37.5</b>

The Proposed Method exhibits superior computational efficiency with the shortest training and testing times across all datasets. For example, on MSMT17, it completes the training in 110.5 s and testing in 40.1 s, outperforming the Dual-stream Model (138.7/51.4 s). This demonstrates its processing speed and scalability efficiency, as shown in Table 16. The Proposed Method outperforms existing Cross-Modality (RGB vs. IR) methods across all datasets. For instance, MSMT17 achieves 94.7/91.6%, exceeding the Dual-stream Model (93.2/89.1%). This shows its strong ability to handle both RGB and IR modalities with higher accuracy, as in Table 17. The proposed method demonstrates lower

Memory Consumption (MB) than existing methods. For example, MSMT17 consumes 230 MB during training and 215 MB during testing, which is more efficient than methods like the Dual-Stream Model (280/265 MB). This shows its memory efficiency while maintaining performance, as in Table 18. The proposed method shows better energy efficiency than existing methods. For example, MSMT17 consumes 170 mJ during training and 160 mJ during testing, which is more energy-efficient than methods like Dual-stream Model (210/200 mJ). This indicates that the proposed method is optimized for lower energy consumption while maintaining effective performance, as in Table 19.

**Table 17. Cross-modality performance (RGB vs. IR) for train and test datasets across existing methods and the proposed method**

Method	MSMT17 (Train/ Test)	Market1501 (Train/ Test)	DukeMTMC (Train/ Test)	CCUHK03-NP (Train/ Test)
VI-PReID [12]	85.2/80.5	88.4/83.1	90.1/84.8	84.3/79.5
MSPA [13]	87.1/82.2	89.3/84.2	91.0/85.5	85.7/80.6
UNIReID [14]	89.4/84.3	91.5/86.7	92.3/87.8	87.4/82.3
M2FNet [16]	90.7/85.8	92.1/87.4	93.2/88.1	88.2/83.4
GCN [17]	84.3/79.9	87.2/81.8	89.0/83.5	82.6/77.9
Deep-MRL [18]	92.4/88.5	93.2/89.1	94.3/90.2	89.7/85.8
CLIP-like Framework [21]	88.3/83.4	90.2/85.3	91.5/86.6	86.2/81.4
AE-Net [22]	89.5/84.9	91.0/86.5	92.0/87.6	86.7/81.9
Dual-stream Model [23]	93.2/89.1	94.5/90.3	95.2/91.4	90.2/86.3
<b>Proposed Method</b>	<b>94.7/91.6</b>	<b>96.2/93.1</b>	<b>97.1/94.5</b>	<b>93.5/89.9</b>

**Table 18. Memory Consumption (MB) for train and test datasets across existing methods and the proposed method**

Method	MSMT17 (Train/ Test)	Market1501 (Train/ Test)	DukeMTMC (Train/ Test)	CCUHK03-NP (Train/ Test)
VI-PReID [12]	220/210	230/215	235/225	210/200
MSPA [13]	210/205	220/210	225/215	205/195
UNIReID [14]	250/240	260/245	265/255	245/235
M2FNet [16]	275/260	285/270	290/275	270/260
GCN [17]	230/220	240/225	245/235	225/215
Deep-MRL [18]	265/250	275/260	280/270	260/250
CLIP-like Framework [21]	240/225	250/235	255/245	235/225
AE-Net [22]	255/240	265/250	270/260	250/240
Dual-stream Model [23]	280/265	290/275	295/285	275/265
<b>Proposed Method</b>	<b>230/215</b>	<b>240/225</b>	<b>245/235</b>	<b>225/215</b>



**Table 19. Energy efficiency (mJ) for train and test datasets across existing methods and the proposed method**

Method	MSMT 17 (Train/ Test)	Market1501 (Train/ Test)	DukeMTMC (Train/ Test)	CCUHK03-NP (Train/ Test)
VI-PReID [12]	160/150	170/160	175/165	160/150
MSPA [13]	150/140	160/150	165/155	150/140
UNIReID [14]	180/170	190/180	195/185	180/170
M2FINet [16]	200/190	210/200	215/205	200/190
GCN [17]	170/160	180/170	185/175	170/160
Deep-MRL [18]	190/180	200/190	205/195	190/180
CLIP-like Framework [21]	180/170	190/180	195/185	180/170
AE-Net [22]	195/185	205/195	210/200	195/185
Dual-stream Model [23]	210/200	220/210	225/215	210/200
<b>Proposed Method</b>	<b>170/160</b>	<b>180/170</b>	<b>185/175</b>	<b>170/160</b>

## 5. Conclusion

The proposed method shows appreciable improvements over a range of criteria when compared to methods now in use on several datasets (MSMT17, Market1501, DukeMTMC, and CCUHK03-NP). Low consumption in the train and test phases results in remarkable energy efficiency, qualities required for practical application, particularly in environments with limited energy availability. Resilient and highly performing in a variety of conditions, the technique also performs remarkably well in terms of Accuracy, Mean Average Precision (mAP), F1-score, CMC, Rank-1, and Top-5 Accuracy. The proposed method proves a great degree of cross-modality performance (RGB against IR), demonstrating its adaptability in many input modalities. This is something

that is rather difficult to achieve in applications applied in the real world. Apart from its remarkable memory consumption, which makes it more suitable for usage in devices with limited resources, it also stands out for better computational efficiency. These results reveal the method's flexibility, effectiveness, and practicality in environments that mirror the real world, in which variables including energy consumption, computational resources, and performance are of main relevance. Apart from defining a new benchmark for future research in this field, the proposed method presents a feasible solution for large-scale, multimodal person re-identification problems. It is a strong candidate for pragmatic deployment in dynamic surroundings since it finds a mix between performance and resource economy.

## References

- [1] He Li et al., "All in One Framework for Multimodal Re-Identification in the Wild," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 17459-17469, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] R. Manikandan et al., "Sequential Pattern Mining on Chemical Bonding Database in the Bioinformatics Field," *AIP Conference Proceedings*, Krishnagiri, India, vol. 2393, no. 1, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Hamza Mukhtar, and Muhammad Usman Ghani Khan, "CMOT: A Cross-Modality Transformer for RGB-D Fusion in Person Re-Identification with Online Learning Capabilities," *Knowledge-Based Systems*, vol. 283, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] S.S. Sivasankari et al., "Classification of Diabetes using Multilayer Perceptron," *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, Ballari, India, pp. 1-5, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Yuvaraj Natarajan, Srihari Kannan, and Sachi Nandan Mohanty, "Survey of Various Statistical Numerical and Machine Learning Ontological Models on Infectious Disease Ontology," *Data Analytics in Bioinformatics: A Machine Learning Perspective*, pp. 431-442, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Keyang Cheng et al., "BAMG: Text-Based Person Re-Identification via Bottlenecks Attention and Masked Graph Modeling," *Proceedings of the Asian Conference on Computer Vision*, pp. 1809-1826, 2024. [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Radjarejesri Shesayar et al., "Nanoscale Molecular Reactions in Microbiological Medicines in Modern Medical Applications," *Green Processing and Synthesis*, vol. 12, no. 1, pp. 1-8, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Gaurav Dhiman et al., "Multi-Modal Active Learning with Deep Reinforcement Learning for Target Feature Extraction in Multi-Media Image Processing Applications," *Multimedia Tools and Applications*, vol. 82, no. 4, pp. 5343-5367, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Di Wu et al., "LRMM: Low Rank Multi-Scale Multi-Modal Fusion for Person Re-Identification based on RGB-NI-TI," *Expert Systems with Applications*, vol. 263, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [10] Xinyu Zhang, Peng Zhang, and Caifeng Shan, "Corruption-Invariant Person Re-Identification via Coarse-to-Fine Feature Alignment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 2, pp. 1084-1097, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Nianchang Huang et al., "Cross-Modality Person Re-Identification via Multi-Task Learning," *Pattern Recognition*, vol. 128, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Samee Ullah Khan et al., "Visual Appearance and Soft Biometrics Fusion for Person Re-Identification using Deep Learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 3, pp. 575-586, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Cuiqun Chen, Mang Ye, and Ding Jiang, "Towards Modality-Agnostic Person Re-Identification with Descriptive Query," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, pp. 15128-15137, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Nianchang Huang et al., "Exploring Modality-Shared Appearance Features and Modality-Invariant Relation Features for Cross-Modality Person Re-Identification," *Pattern Recognition*, vol. 135, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Jianan Liu, Jian Liu, and Qiang Zhang, "M2FINet: Modality-Specific and Modality-Shared Features Interaction Network for RGB-IR Person Re-Identification," *Computer Vision and Image Understanding*, vol. 232, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Guang Han et al., "Text-to-Image Person Re-identification Based on Multimodal Graph Convolutional Network," *IEEE Transactions on Multimedia*, vol. 26, pp. 6025-6036, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Suncheng Xiang et al., "Deep Multimodal Representation Learning for Generalizable Person Re-Identification," *Machine Learning*, vol. 113, no. 4, pp. 1921-1939, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Xiangtian Zheng et al., "Multi-Modal Person Re-Identification based on Transformer Relational Regularization," *Information Fusion*, vol. 103, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Cosimo Patruno et al., "Multimodal People Re-identification using 3D Skeleton, Depth and Color Information," *IEEE Access*, vol. 12, pp. 174689-174704, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Meng Zhang, Rujie Liu, and Abe Narishige, "Face Helps Person Re-Identification: Multi-modality Person Re-Identification Based on Vision-Language Models," *2024 IEEE International Joint Conference on Biometrics (IJCB)*, Buffalo, NY, USA, pp. 1-10, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Yongkang Ding et al., "Attention-Enhanced Multimodal Feature Fusion Network for Clothes-Changing Person Re-Identification," *Complex & Intelligent Systems*, vol. 11, no. 1, pp. 1-15, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Qingshan Chen et al., "MSIF: Multi-Spectrum Image Fusion Method for Cross-Modality Person Re-Identification," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 2, pp. 647-665, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Jiaxuan Li et al., "Multimodal Feature Hierarchical Fusion for Text-Image Person Re-Identification," *In Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, Urumqi, China, pp. 468-481, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]