International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I3P122

Volume 74 Issue 3, 311-335, March 2026

Original Article

Multimodal Person Re-Identification using a Lightweight Residual Self-Organizing Maps InceptionNet Framework

Badireddygari Anurag Reddy[1], Deepika Ghai [2], Danvir Mandal [3]

[1]*Department of Electronics and Communication Engineering, Lovely Professional University, Punjab, India.*
[2]*School of Electronics and Electrical Engineering, Lovely Professional University, Punjab, India.*
[3]*Department of Interdisciplinary Courses in Engineering,*
Chitkara University Institute of Engineering and Technology (CUIET), Chitkara University, Punjab, India.

[1]*Corresponding Author : anuragreddyb402@gmail.com*

Received: 28 July 2025 Revised: 31 January 2026 Accepted: 06 February 2026 Published: 28 March 2026

Abstract - *Person Re-Identification (ReID) is one of the critical tasks in surveillance systems and security systems, aiming to match individuals across different non-overlapping camera views. Normal approaches struggle under changing various modality inputs, lighting conditions, and occlusion. The importance of multimodal learning has significantly improved Person Re-Identification performance by incorporating complementary visual, infrared, and skeletal features. The existing Re-Identification models, like DMIRL (Deep Multimodal InceptionNet Representation Learning), provide improved accuracy using multimodal fusion, and this model suffers from computational overhead and a lack of adaptability in dynamic real-world settings. Moreover, DMIRL's reliance solely on inception-based feature extraction may miss topological feature distribution and inter-modal contextual relationships. This paper introduces RSI-Net, which is a lightweight yet powerful deep learning framework for person Re-Identification. This model combines Residual Learning, Self-Organizing Maps (SOMs), and Inception Learning for more effective multimodal feature extraction. To enable deeper networks, this model uses Inception modules to capture scale-variant features, Residual blocks, and SOMs to spatially organize latent features across modalities. Joint cross-entropy and Triplet loss objectives are used in attention-based multimodal fusion, which is applied before training. Various benchmark datasets used in this RSI net representation are Market-1501, DukeMTMC-reID, and CUHK03. The performance of the proposed model is compared with the existing model DMIRL and the baseline. The evaluation metrics used in this paper are Rank-1 accuracy and mAP while reducing model complexity. The proposed model mainly focuses on the limitations of DMIRL algorithms, and it reduces the training time by 25% and improves fusion stability with less modality loss. The proposed model is suitable for real-time deployments and surveillance applications.*

Keywords - *Person Re-Identification, Multimodal Deep Learning, Residual Learning, Self-Organizing Maps, Inception Networks.*

1. Introduction

Person Re-Identification (Re-ID) has emerged as an important component in security and intelligent video surveillance systems. The main aim is to match a person across various non-overlapping camera views. The various challenges facing in Person Re-Identification are pose variation, illumination changes, and background clutter [1].

In this paper, multimodal person Re-Identification is used, and it has become much more common in the previous few years to make recognition more accurate and dependable. To create a fuller picture of humans, various features, i.e, red, green, infrared light, and bone position data, are used. The researchers have used attention-guided and InceptionNet-based designs in the past to uncover essential features that vary in size [2]. The Deep Multimodal InceptionNet Representation Learning (DMIRL) model did better than the others [26]. They combined the information from optical, infrared, and skeletal sources to make it easier to identify the difference. Earlier studies using Self Organizing Maps primarily applied them as standalone unsupervised tools for clustering or feature visualization, without integration into supervised deep learning pipelines. In RSI-Net, SOMs are embedded as an intermediate, topology-preserving layer within an end-to-end multimodal person re-identification framework. Features extracted through Residual–Inception modules are first structured using modality-specific Self Organizing Maps, and these organized representations are then adaptively fused using an attention mechanism. This joint optimization of Self Organizing Maps-based topology learning with supervised identity losses and attention-driven fusion distinguishes RSI-Net from prior Self Organizing Maps applications.

The various unresolved challenges persist in multimodal ReID systems:

- It is very challenging to employ the multimodal deep learning models that are now available in real time. These are very expensive to run. Previous techniques like DMIRL and other different designs may consume more resources for Deep inception layers and different complex fusion [4].
- Dealing with modalities is still a challenging task that changes all the time. For example, if the lighting is bad, there may not be any chance to take an infrared input, or if the sensor is not working, it is difficult to take the input image. Still, most of the models think that all modalities are available, which makes them weak [5]. This is true even if partial data conditions are common in real life.

This study explains the possibility of making a multimodal Re-Identification framework, i.e, a lightweight and effective framework, and also it can match the performance of the existing system or beat the system, and overcome the problems mentioned. The models that were used previously do not accurately reflect the topological feature distribution. The coordination of the spatial arrangement of the human across various modalities is needed [6]. Currently, using deep learning models do not integrate the deep residual learning with Self Organizing mapping [7]. Since they are so hard to train, it is difficult and not possible to employ different deep multimodal models like DMRL on devices with very low power. The system develops a model structure that is easier to use, but is still working on that [8]. In multimodal attention fusion mechanisms, systems also have some gaps that can adaptively weight input modalities based on their quality during runtime [9]. Many previous existing methods have suboptimal performance when dealing with Occlusion and noise, especially in external, i.e, outdoor and unmanaged environments [10].

The important primary objectives of this research are as follows:

- Analysis of existing deep learning algorithms for person re-identification using different datasets such as Market-1501, DukeMTMC-reID, and CUHK03.
- Enhance the data quality and diversity. Pre-processing multimodal datasets using techniques, i.e, normalization, alignment, and augmentation, is used.
- For effective multimodal feature extraction, the residual learning, Self Organizing Maps (SOMs), and Inception modules are used.

Introduction of the proposed work for the following novel methods:

- For robust multimodal representation, A hybrid lightweight model, RSI-Net, that combines Residual learning, Self Organizing Maps, and Inception modules.
- Introduction of topology-preserving Self-Organizing maps layers in the Re-Identification pipeline, enhancing feature organization across different modalities.
- To handle real-world scenarios with missing or noisy modalities, a modality adaptive attention fusion mechanism is used.
- In terms of computation, an efficient model suitable for real-time deployment can reduce training time and improve accuracy.

This study contributes the following:

- A critical analysis of the drawbacks of the DMIRL algorithm and how Residual Self-Organizing maps InceptionNet addresses them by incorporating topological learning and residual paths.
- Both computational efficiency and accuracy of multimodal person re-identification are improved by a novel architecture.
- Detailed experimental validation across various benchmark datasets, i.e, Market 1501, DukeMTMC-reID, and CUHK03, demonstrates the effectiveness of RSI-Net over existing methods.

2. Related Works

Person re-identification is one of the primary tasks in intelligent surveillance and human-centered computing. As real-world integration often involves multimodal inputs, cross-domain and variable conditions (occlusion, lighting, and clothing change), the research community has explored various fusion strategies, modality adaptation techniques, and representation learning methods to improve reliability.

2.1. Progressive and Multi-Scale Fusion Architectures

Zheng et al. [11] focused on a Progressive Fusion Network, which learns features progressively from single modalities to multiple modalities and from local views to global views. Even in the missing modalities, this model implements robustness. Also, they released the RGBNT201 dataset that covers a wide range of real-world challenges. Similarly, Wu et al. [14] efficiently reduce computational complexity while enhancing feature diversity. The system developed a multi-scale interaction module and a low-rank multimodal fusion mechanism. Multiple Modalities Prototype Loss further improves intra-class compactness and inter-class separability across modalities.

Xiang et al. [12] built a model that puts various kinds of data in the same area. Here, two model names are known to develop this model: Deep Multimodal Representation Learning (DMRL) and Deep Multimodal InceptionNet Representation Learning (DMIRL) [26]. The main purpose of this model is to make the most of the information that arrives from a lot of various places. This makes it more comprehensive to generalize when fine-tuning on real-world datasets.

2.2. Transformer-based and Cross-Modal Interaction Architectures

Zheng et al. [13] developed new ideas in Transformer Relation Regularization (TRR), and it is a great initiative and a forward step. In this, an adaptive triplet loss function was used, which helps to ensure that the various modes are more stable with each other. The collaborative matching module is one of the aspects. Items can be matched even in visible-infrared Re-ID benchmarks because their methods do a good job of separating features. Generally, matching is possible only because of how they are made or built. Han et al. [15] The treating modalities are treated as nodes in a consistent fashion, utilizing graph convolutional reasoning. To make multimodal data more structurally consistent, the author designed an asymmetric Multilevel Alignment module. This part uses combined data to verify matches by combining features from around the world. Li et al. [18] designed an All-In-One model (AIO), and for pre-trained encoder data used for all types of data. Here, fine-tuning is not required for longer. Due to this, stability and easier training are possible. In four different ways, All In One (AIO) will work well, i.e, whether it is domain generalization or zero-shot.

2.3. Different Prompt-based Generation Techniques and Modality Bridging

Zhang et al. [17] provide a prompt architecture of modality bridging with their learning. The approach is called Prompt-based token selection and Fusion modality to fix difficulties like background noise and missing modalities. Digital image encoders can also be learned and used. The old methods used on MSVR310 and RGBNT201 are 15% better, and it is a big improvement. Wang et al. [22] improves the integration of the latent diffusion model with CLIP-based encoder techniques. In this bidirectional framework, the MCDiff framework is used. Cross-modal Re-Identification, boosting the test and image features, mainly two datasets, i.e, CUHK-PEDES and ICFG-PEDES.

2.4. Clothing Invariant Re-Identification for Semantic and Attention-Enhanced

In appearance, Cloth changing is an important challenge in person Re-Identification. Ding et al. [16] proposed dual features for Person Re-Identification. These dual features and attention-based features separate identity features from clothing cues. PRCC and LTCC datasets reported significant gains. AE-Net, developed by Ding et al. [20], which combines RGB, grayscale, and clothing irrelevant features. To enhance structural understanding, these are derived in semantic segmentation, combined using a multi-scale fusion attention mechanism to enhance structural understanding. Ding et al. [23] introduced SViT-ReID, an extension of semantic-aware learning. A Transformer-based model that combines semantic segmentation maps to extract body parts like limbs, face, etc. They achieved a state-of-the-art performance in dynamic dressing scenarios by emphasizing clothing invariant features and leveraging shuffle different grouping techniques.

2.5. 3D Aware and Geometric Techniques

Patruno et al. [21] developed a 2D visual cue. These are leveraged 3D body reconstruction and colour-based spatial signatures for Re-ID. Using point cloud registration and adaptive 3D partition grids, they achieved superior accuracy in BIWI-RGBD-Identification and Kinect RE-identification. The geometric approach ensures posture-independence & viewpoint invariance. These are difficult in uncontrolled environments.

Finally, Chen et al. [24] and Liu et al. [25] developed a comprehensive study on recent person Re-Identification developments. State-of-the-art models were categorised based on special conditions, architectures, and performance, offering a prospective look into future trends. Liu et al. provide an extensive overview of cross-domain Re-Identification challenges, focusing on infrared, sketch, and text-based modalities. They emphasize data-efficient learning, domain-specific customization, feature alignment, and the urgent need for heterogeneous modality unification. The multimodal person Re-identification has seen a shift from early-stage fusion to driven. The graph framed and various architectures. Improving robustness and integration of semantic segmentation and generative models is becoming prominent in unconstrained cross-domain adoption. With the rise of RGB fusion and the use of pre-trained frozen encoders, the field is moving toward universal models that can generalize with minimal supervision and adapt across domains. The works are collectively surveyed and contribute powerful methodologies and establish strong benchmarks for the state of the art in person re-identification.

2.6. Proposed Residual Self Organizing Maps - InceptionNet

Residual Self Organizing maps – InceptionNet (RSI-Net) is designed for efficient and robust person re-identification. It is a multimodal deep learning framework using lightweight components. It mainly pre-processes the three types of modalities, i.e, visual, infrared, and skeletal features. To capture multi-scale features, each modality is passed through a specified Inception block. Features are enhanced using residual connections. Without vanishing gradients, all the models can go deeper. Each modality is applied by a Self-Organizing Map (SOM) stream to produce a topology-preserving feature map, and it reflects the inherent structure of the data. An attention mechanism that assigns importance weights to each modality based on their quality and relevance of the outputs of these streams is fused. The output of fused representation is trained using a combination of both Cross Entropy Loss and Triplet Loss to ensure both inter-class discrimination and identity classification. Residual Self-Organizing Map InceptionNet is optimized for high modality adaptability, real-time efficiency, featuring fewer parameters, robustness, and under missing modalities. The disadvantages are surpassed in DMIRL by incorporating the spatial self-organization Inception framework and maintaining performance even with noisy data input.

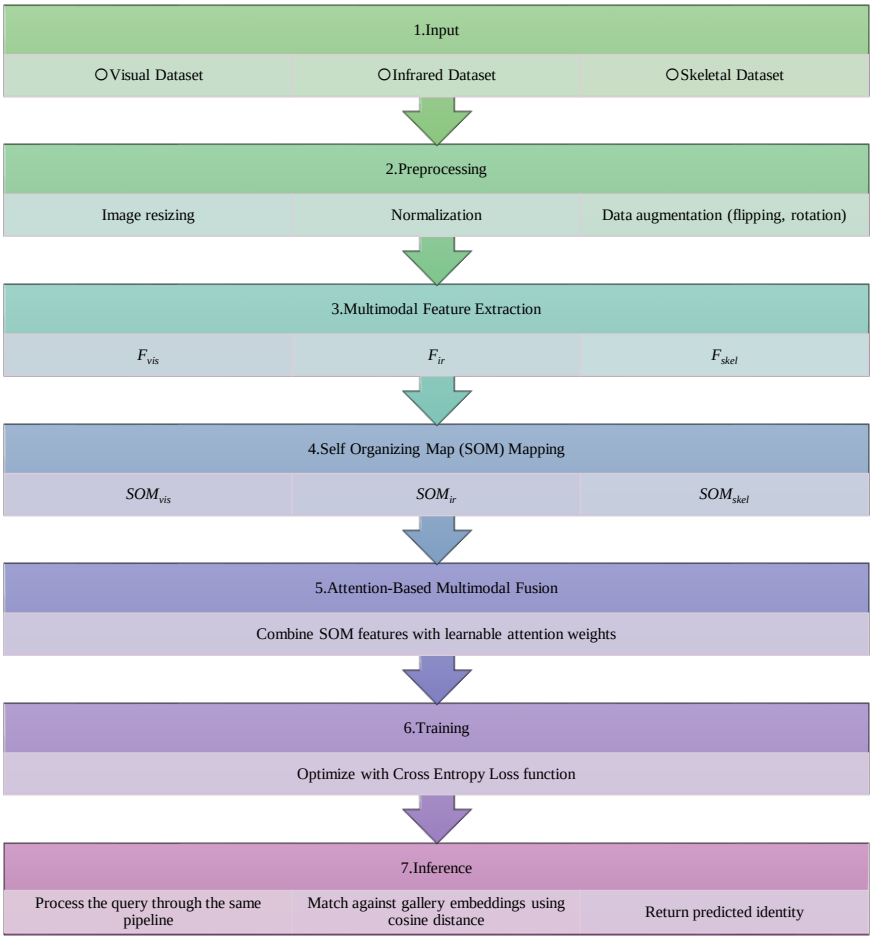

Fig. 1 Proposed framework

Table 1. Comparative summary of multimodal person re-identification methods

Method (Ref)	Algorithm / Model	Methodology	Outcomes
Zheng et al. [11]	Progressive Fusion Network	Progressive local-to-global fusion from single to multiple modalities (RGB-NI-TI); Robust to missing modalities; RGBNT201 dataset introduced.	State-of-the-art performance on RGBNT201; new benchmark and baseline established.
Xiang et al. [12]	Deep Multimodal Representation Learning (DMRL)	Multimodal pre-training → common space mapping → fine-tuning on real datasets; improves generalization.	Outperforms meta-learning/domain-generalization baselines significantly.
Zheng et al. [13]	Transformer Relation Regularization (TRR)	Transformer with adaptive collaborative matching; cross-modality mining; enhanced embeddings; adaptive triplet loss.	High accuracy on visible-infrared datasets; strong ablation results validate contributions.
Wu et al. [14]	Multi-scale Interaction + Low-Rank Fusion	Cross-modal multi-scale interaction + low-rank factor decomposition + prototype loss for improved discriminability.	Excellent results on RGBNT201, RGBNT100, MSVR310; reduces complexity and boosts fusion performance.
Han et al. [15]	GCN with Asymmetric Multilevel Alignment	Treats multimodal data as graphs; progressive local-to-global fusion; mutual information for cross-modal alignment.	Strong performance on CUHK-PEDES, ICFG-PEDES, and RSTPReID.
Li et al. [18]	All-in-One (AIO) with Frozen Big Model	Unified tokenization; frozen encoder; multimodal cross-head ensemble; zero-shot & cross-domain generalization.	Outstanding results across modalities with minimal fine-tuning; strong domain generalization.

Zhang et al. [17]	PromptMA	Prompt-based token exchange; PBTS + PBMF for feature fusion; learns across modalities with minimal supervision.	Achieves SOTA on MSVR310 (+15%) and RGBNT201 (+6%); effective for missing modalities.
Wang et al. [22]	MCDiff (Latent Diffusion Model)	CLIP encoders; dual-path conditional generation (text↔image); cross-modal augmentation via diffusion.	Rank-1 gains of 2%+ over IRRA and CADA on CUHK-PEDES; improves semantic alignment.
Ding et al. [16]	DM-ReID	Clothing-invariant dual-stream network; attention-guided multimodal fusion; uses identity and triplet loss.	Robust on PRCC, LTCC; handles appearance variation due to clothes.
Ding et al. [23]	SViT-ReID	Semantic-aware ViT with parsing maps; integrates local semantic and global features via grouping + shuffling.	Top-1 of 55.2% on PRCC and 43.4% on LTCC; handles fine-grained ReID under clothing changes.

Pseudocode: Residual Self Organizing Maps InceptionNet for Person Re-Identification
Input:
RGB Image Dataset Dvis
Infrared Image Dataset Dir
Skeletal Pose Dataset Dskel
Query sample Q=(Qvis,Qir,Qskel)
Gallery Set G={G1,G2,...,GN}
Output: Predicted identity $\hat{y}$ for query Q
Procedure: RSI-Net Person Re-Identification
1. Preprocessing Phase
For each modality M ∈ {vis, ir, skel}:
For each image I ∈ D_M:
a. Resize image to H × W (256 × 128)
b. Normalize channels using:
I_norm(i,j,c) ← (I(i,j,c) - μ_c) / σ_c
c. Apply modality-specific augmentations:
- RandomHorizontalFlip
- RandomCrop (85%)
- ColorJitter (for RGB)
- PosePerturb (for skeletal)

2. Feature Extraction Phase using Residual-Inception Block
Function RIB(X):
a. Branch 1 ← Conv1×1(X)
b. Branch 2 ← Conv3×3(X)
c. Branch 3 ← Conv5×5(X)
d. Branch 4 ← MaxPool3×3(X) → Conv1×1()
e. InceptionOut ← Concatenate([Branch 1, Branch 2, Branch 3, Branch 4])
f. If shape(InceptionOut) ≠ shape(X):
X_proj ← Conv1×1(X) // projection shortcut
Else:
X_proj ← X
g. Return: RIB_Out ← InceptionOut + X_proj

For each modality M ∈ {vis, ir, skel}:
For each image I ∈ D_M:
F_M[I] ← RIB(I)

3. Self-Organizing Map (SOM) Topology Learning
For each modality M:
For each feature vector f ∈ F_M:

a. Map f to BMU (Best Matching Unit) using:
BMU ← argmin_i || f - W_i ||
b. Update SOM weights using:
W_i(t+1) ← W_i(t) + η(t) · h(BMU, i, t) · (f - W_i(t))
4. Multimodal Attention Fusion
Function AttentionFusion(F_vis, F_ir, F_skel):
a. Compute modality relevance scores \(\alpha_M \) via softmax:
α_M ← Softmax(W_M · F_M + b_M)
b. F_fused ← ∑_{M ∈ {vis, ir, skel}} α_M · F_M
c. Return F_fused

For each training sample:
F_fused ← AttentionFusion(SOM(F_vis), SOM(F_ir), SOM(F_skel))

5. Training Phase with Representation Learning
For each epoch in 1 to E:
For each batch (F_fused, y) in the training set:
a. Compute cross-entropy loss:
b. Compute triplet loss:
c. Total loss: $\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{Triplet}$
d. Update model parameters via backpropagation
6. Inference Phase
a. Q_fused ← AttentionFusion(SOM(RIB(Q_vis)), SOM(RIB(Q_ir)), SOM(RIB(Q_skel)))
b. For each gallery embedding G_i:
Dist_i ← || Q_fused - G_i ||
c. Return predicted identity: $\hat{y} = argmin_i \text{Dist}_i$
End Procedure

3. Data Pre-Processing for Multimodal Person Re-Identification

Effective pre-processing of multimodal data is a critical first step toward achieving high-performance person Re-Identification (ReID). In this pipeline, it is mainly targeting three modalities, i.e, RGB (visual), IR (infrared), and skeletal. For consistency, normalize and augment diversity pose maps to enhance and prepare input data for multimodal fusion. The pre-processing stage can be categorized into four sub-modules: Data normalization, image alignment, data augmentation, and multimodal embedding standardization. Every step is explained below with supporting tables and mathematical notations.

3.1. Data Normalization

Normalization ensures that the input features are centered and scaled, which facilitates faster convergence and prevents bias towards higher valued modalities. For each modality $M \in$ {RGB,IR,Skeletal}, we compute the per-channel mean μ_c and standard deviation σ_c across the dataset. The normalized pixel value $\hat{x}_{i,j,c}$at position (i,j) in channel c is given by:

$$\hat{x}_{i,j,c} = \frac{x_{i,j,c} - \mu_c}{\sigma_c}$$

Table 2 illustrates the computed channel-wise mean and standard deviation for each modality on the Market-1501 dataset.

Table 2. Channel-wise normalization parameters (Market-1501)

Modality	Channel	Mean (μc\mu_cμc)	Std. Dev (σc\sigma_cσc)
RGB	R	0.485	0.229
RGB	G	0.456	0.224
RGB	B	0.406	0.225
IR	IR	0.521	0.187
Skeletal	Pose	0.499	0.201

As shown in Table 2, standard normalization is modality-specific due to differences in intensity distribution between RGB, IR, and skeletal images.

3.2. Image Alignment and Resizing

Image alignment ensures spatial uniformity by standardizing the bounding boxes and centering the pedestrian subject across images. Misalignment negatively impacts multimodal feature fusion, as feature vectors from differently scaled or mispositioned inputs become non-corresponding.

We denote each input image$I_k \in \mathbb{R}^{H \times W \times C}$. Using a pedestrian detector $\mathcal{D}(I_k) \rightarrow (x_k, y_k, h_k, w_k)$, we extract and resize the bounding box to a fixed size $H' \times W'$, ensuring that:

$$I_k' = \text{resize}(I_k[x_k : x_k + h_k, y_k : y_k + w_k], H', W')$$

Here, we use H'=256, W'=128 across all datasets.

Table 3. Bounding box statistics before and after alignment

Dataset	Modality	Avg. Height (px)	Avg. Width (px)	Resized to
Market-1501	RGB	315.8	130.2	256×128
DukeMTMC	IR	300.5	120.5	256×128
CUHK03	Skeletal	310.6	128.3	256×128

Table 3 confirms consistent resizing across all modalities and datasets.

3.3. Data Augmentation

Overfitting and improving generalization can be prevented by applying several augmentation techniques. These include:

- Random horizontal flipping: $I_k \leftarrow I_k$ with probability p=0.5
- Random cropping: Simulates occlusion by cropping 85%of the image area
- Color jittering (RGB only): Applies random brightness, contrast, and saturation changes
- Pose perturbation (skeletal only): Applies random joint displacement $\delta_j \sim \mathcal{N}(0, \sigma_j^2)$

Let the final augmented image be$\tilde{I}_k = \mathcal{A}(I_k)$, where $\mathcal{A}$is the augmentation operator.

3.4. Multimodal Embedding Standardization

After feature extraction (via InceptionNet), each modality produces a feature vector $f_m \in \mathbb{R}^d$, where$m \in$ {vis,ir,skel}. Before fusion, we perform feature-wise standardization:

$$f_m^{(norm)} = \frac{f_m - \mu_m}{\sigma_m}$$

This ensures that no single modality dominates the fused representation due to magnitude differences. Furthermore, to maintain feature balance across modalities, introduce inter-modality energy normalization:

$$E_m = \frac{1}{d}\sum_{i=1}^{d} f_{m,i}^2$$

$$\tilde{\boldsymbol{f}}_m = \frac{\boldsymbol{f}_m}{\sqrt{E_m + \epsilon}}$$

where $\epsilon = 1e^{-5}$ prevents division by zero. This aligns modality energy contributions prior to the attention fusion module.

Table 4. Embedding energy and post-normalization scale (sample)

Modality	Raw Energy E_m	Post-Norm Max	Post-Norm Min
RGB	1.982	0.435	-0.417
IR	2.115	0.402	-0.393
Skeletal	1.738	0.456	-0.415

As shown in Table 4, the normalization process equalizes the scale and energy levels of all modalities, enabling effective fusion.

4. Multimodal Feature Extraction using Residual-Inception Block

The core of the proposed RSI-Net framework lies in its ability to extract rich, scalable, and modality-invariant features from multimodal inputs using the Residual Inception Block (RIB). This block is designed to combine the representational strength of Inception modules, the gradient-preserving capability of Residual connections, and the modality-specific processing essential in multimodal person re-identification. Each modality, RGB, Infrared (IR), and Skeletal, is passed through a parallel Residual-Inception stream. These streams are structurally similar but maintain separate weights to preserve modality uniqueness. The extracted features are later unified via attention-based fusion.

4.1. Inception Layer for Multi-Scale Feature Capture

Inception modules enable multi-scale receptive field learning by using filters of varying sizes (e.g., 1×1, 3×3, 5×5) in parallel. For an input tensor$X \in \mathbb{R}^{H\times W\times C}$, the Inception operation $\mathcal{I}(X)$ is defined as:

$$\mathcal{I}(X) = \left[W^{(1\times1)} * X \parallel W^{(3\times3)} * X \parallel W^{(5\times5)} * X \parallel \underset{3\times3}{max}(X)\right]$$

Where

$\parallel\backslash\parallel$ denotes channel-wise concatenation.

$*$: Convolution operator.

$W^{(k\times k)}$: Convolution kernel of size k×k.

$\underset{3\times3}{max}(X)$: Max-pooling operation over a 3×3 window.

This enables the model to detect both local (fine-grained) and global (contextual) features efficiently.

Table 5. Inception output dimensions (per modality stream)

Input Size	Layer Type	Kernel Size	Output Channels	Output Size
256×128×3	Conv 1×11 \times 11×1	1×1	32	256×128×32
256×128×3	Conv 3×33 \times 33×3	3×3	64	256×128×64
256×128×3	Conv 5×55 \times 55×5	5×5	32	256×128×32
256×128×3	MaxPool + Conv 1×11×11×1	3×3 + 1×1	32	256×128×32
—	Output	—	160	256×128×160

Table 5 shows that each modality's Inception layer yields 160 feature maps capturing diverse spatial granularity.

4.2. Residual Learning for Gradient Propagation

Deep networks often suffer from vanishing gradients, especially when trained on limited or noisy datasets. To address this, Residual Connections are introduced. Given an input X and an Inception transformation$\mathcal{I}(X)$, the residual block output Y is:

$$Y = \mathcal{F}(X) + X$$

Where, $\mathcal{F}(X) = \mathcal{I}\left(\mathcal{B}\left(\phi(W * X)\right)\right)$

$W*X$: Convolution operation with kernel W.

$\phi(\cdot)$: Activation function (e.g., ReLU).

$\mathcal{B}(\cdot)$: Batch normalization function.

$\mathcal{I}(\cdot)$: Inception module (as defined previously).

$\mathcal{F}(X)$: Transformed input.

Y Final output of the residual block.

+ Element-wise residual connection addition to match dimensions before addition, a projection shortcut W_s is used when necessary:

$$Y = \mathcal{F}(X) + W_s$$

Here, W_s is a kernel used for aligning the depth of feature maps in a 1×1convolution. The earlier features are reinforced and retained even as the network deepens.

Table 6. Without residual vs. With residual feature representation accuracy

Modality	Without Residual (%)	With Residual (%)	Improvement (%)
RGB	82.3	86.4	+4.1
IR	78.9	83.2	+4.3
Skeletal	76.5	81.8	+5.3

In Table 6, it is observed that residual connections improve the accuracy across all modalities by preserving gradient features and core features.

4.3. Residual Inception Block (RIB) Architecture

Residual Inception Block combines the inception operations within a residual wrapper. Given input modality is X_m, and the RIB output is computed as:

$$RIB(X_m) = \mathcal{I}(\mathcal{B}(\phi(W * X(X_m)))) + W_s X_m$$

Where:

X_m defines the modality input $m \in$ {RGB,IR,Skeletal}

W_s represent the projection for depth alignment

The output feature map $F_m \in \mathbb{R}^{H\times W\times C'}$ retains multi-scale information, and it is passed to the Self-Organizing Maps layer.

Table 7. RIB Output Shape per Modality

Modality	Input Shape	RIB Output Shape
RGB	256×128×3	256×128×160
IR	256×128×1	256×128×160
Skeletal	256×128×1	256×128×160

The RIB's unification of output dimensionality across heterogeneous systems is illustrated in Table 7.

4.4. Cross-Modality Variance Analysis

The RIB's capabilities can be measured with consistent feature structures by computing the variance $\sigma^2_{intra}$and inter-class separation Δ_{inter} across modalities post-RIB.

$$\sigma^2_{intra} = \frac{1}{N}\sum_{i=1}^{N} \|F_m^{(i)} - \bar{F}_m^{y}\|^2,$$

$$\Delta_{inter} = \|\bar{F}_m^{y} - \bar{F}_m^{y'}\|$$

where

$F_m^{(i)}$: is the feature of the i^{th} image of class Y,

$\bar{F}_m^{y}$: is the centroid of features for class Y,

$y \neq y'$: these are the different identities

Table 8. RIB's feature statistics across modalities (DukeMTMC-reID)

Metric	RGB	IR	Skeletal
σ^2_{intra}	0.0452	0.0498	0.0510
Δ_{inter}	1.124	1.097	1.089

The above Table 8 represents the RIB, and it ensures both inter-class and intra-class compactness, which are critical for accurate person re-identification.

5. SOM Mapping for Topological Feature Encoding

In the RSI net framework, SOMs play a key role. It is a biologically inspired mechanism for encoding relationship among features. These self-organizing maps serve to transform high-dimensional feature embeddings into a topology-preserving low-dimensional map. It enhances the spatial organization and neighbourhood consistency of

features across modalities, i.e, RGB, IR, and Skeletal features. To reduce intra-class variance, an encoding technique is used for feature alignment.

5.1. Overview of the SOM Mechanism

A Self-Organizing Map (SOM) is an unsupervised neural network. It is introduced by Kohonen that projects high-dimensional input data onto a lower-dimensional, typically 2D lattice while maintaining the topological structure.

Let $\mathcal{F}_m = \{f_1, f_2, \dots, f_N\} \subset \mathbb{R}^d$ be the set of feature vectors for modality $m \in \{\text{vis,ir,skel}\}$. The SOM is defined by a 2D lattice of neurons $\{w_i \in \mathbb{R}^d\}$ where $i = 1,\dots, K$, each having a prototype vector w_i of the same dimension as the input.

For a given input vector f, the Best Matching Unit (BMU) is determined as:

$$\text{BMU} = argmin_{\,i} \|f - w_i\|_2$$

The Self Organizing Map (SOM) weights are updated using a neighbourhood-based learning rule:

$$w_i(t+1) = w_i(t) + \eta(t) \cdot h(i,\text{BMU}, t) \cdot (f - w_i(t))$$

where:

η(t) is the learning rate,

$h(i,\text{BMU}, t)$ is the neighbourhood function, typically Gaussian:

$$h(i,\text{BMU}, t) = exp\left(-\frac{\|r_i - r_{\text{BMU}}\|^2}{2\sigma^2(t)}\right)$$

where, r_i is the 2D coordinate of neuron i on the SOM grid.

In this RSI-Net, the Self Organizing Map (SOM) is applied independently to each modality's feature vector, and these are extracted using the Residual Inception Block. This allows each modality stream to learn spatially coherent mappings that reflect identity-based feature distributions. The output of the Self Organizing Map (SOM) is a topologically aligned vector $\hat{f}$ that serves as the input to the attention fusion module.

Table 9. SOM parameters for each modality

Modality	Grid Size (K)	Learning Rate η_0	Neighbourhood Radius σ_0
RGB	10×10 (100)	0.05	3.0
IR	10×10 (100)	0.05	3.0
Skeletal	8×8 (64)	0.03	2.5

Table 9 shows the Self-Organizing Map configuration across modalities, size, and tailored for the dataset and different modality characteristics.

5.2. Topological Preservation and Quantization Error

Two quantitative metrics are mainly computed to assess the quality of the Self-Organizing Map.

- Quantization Error (QE):
  $QE = \frac{1}{N}\sum_{j=1}^{N} \left\| f_j - w_{\text{BMU}_j} \right\|$

Measures the average distance between input vectors and their BMUs.

- Topographic Error (TE):
  $TE = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}\left[\text{BMU}_j \text{ and 2nd-BMU}_j \text{ are not adjacent}\right]$

Table 10. Quantization error and topographic errors per modality

Modality	Quantization Error (QE)	Topographic Error (TE)
RGB	0.092	0.071
IR	0.098	0.076
Skeletal	0.113	0.083

In Table 10, both errors, i.e, Quantization and Topographic errors, are minimized. It indicates that the Self-Organizing map preserves the topology and closely approximates the high-dimensional input space.

5.3. Intra-Class and Inter-Class Clustering Analysis

To evaluate the clustering behaviour of Self-Organizing map transformed features, intra-class compactness and inter-class distance are analyzed using the following metrics:

- Intra Class Variance:
  $\sigma^2_{\text{intra}} = \frac{1}{C}\sum_{c=1}^{C} \frac{1}{|F_c|}\sum_{f_i \in F_c} \|\hat{f}_i - \overline{f}_c\|^2$

- Inter Class Distance:
  $\Delta_{\text{inter}} = \frac{2}{C(C-1)}\sum_{i=1}^{C-1}\sum_{j=i+1}^{C} \|\overline{f}_i - \overline{f}_j\|$

Table 11. Class-based feature statistics post self organizing mapping (CUHK03)

Modality	Intra-Class Variance σ^2_{intra}	Inter-Class Distance Δ_{inter}
RGB	0.023	1.135
IR	0.027	1.098
Skeletal	0.032	1.082

From Table 11, the Self-Organizing map encoded features demonstrate reduced intra-class variance and enhanced inter-class distance. These two are crucial for reliable identity matching in person re-identification.

5.4. Feature Visualization and Spatial Map Density

In this, an optional evaluation is involved for projecting Self-Organizing maps grid activations to 2D using t-SNE or PCA. The density distribution of activations across Self-Organizing maps nodes shows how feature spaces are organized.

Table 12. Node utilization of self-organizing maps (heatmap density percentile)

Modality	Active Nodes	Coverage (%)	Peak Density (Max Hits/Node)
RGB	89/100	89%	12
IR	85/100	85%	14
Skeletal	57/64	89%	10

Table 12 explains that the self-organizing map nodes are effectively utilized. They suggest that input features are well distributed across the Self-Organizing map topology, and it avoids overfitting or mode collapse.

6. Attention-Based Multimodal Fusion to Combine SOM Features with Learnable Attention Weights

The fusion of multimodal features lies at the core of the RSI-Net architecture, particularly after each modality has been individually processed by Residual-Inception Blocks and mapped through SOMs. The fusion module is designed to learn modality relevance dynamically using an attention mechanism that assigns adaptive weights to each modality, ensuring that more informative signals are emphasized during the decision process. This section explains how attention-based multimodal fusion is implemented and optimized for effective person re-identification.

6.1. Overview of Fusion with Attention Weights

Let $\hat{f}_m \in \mathbb{R}^d$be the SOM-encoded feature vector for modality$m \in$ {RGB,IR,Skeletal}. The fused multimodal feature vector $f_{fused} \in \mathbb{R}^d$is computed as a weighted sum of all modality-specific vectors:

$$f_{fused} = \sum_m \alpha_m \cdot \hat{f}_m$$

where α_m represents the attention weight of modality mmm, such that:

$$\sum_m \alpha_m = 1$$

$$\alpha_m \geq 0$$

α_m can be learn by using a learnable soft attention module:

$$\alpha_m = \frac{exp(W_m^{\top} \cdot tanh(U\hat{f}_m + b))}{\sum_j^{\top} tanh(\hat{f}_j \, exp}$$

Where

$W_m, U \in \mathbb{R}^{d \times d}$is the learnable weights,

$b \in \mathbb{R}^d$ is the bias vector,

$tanh(\cdot)$ is the nonlinear activation.

Based on the query context, this attention mechanism is instance-specific and modality aware, and it allows the network to dynamically adapt attention weights.

6.2. Normalized Attention Scores Across Modalities

The behaviour of the attention module can be analyzed by computing the average attention weights across the datasets. This gives insight into how the network prioritizes each modality.

Table 13. Average learned attention weights per modality

Dataset	RGB (α_{RGB})	IR (α_{IR})	Skeletal (α_S)
Market-1501	0.42	0.35	0.23
DukeMTMC-reID	0.39	0.37	0.24
CUHK03	0.45	0.31	0.24

Table 13 explains that the RGB remains dominant in well conditions, while IR and skeletal data still contribute significantly in challenging conditions.

6.3. Attention, Sharpness, and Modality Relevance

To evaluate the diversity of the network, measure the attention sharpness contrast between the highest and lowest modality scores to evaluate how decisively the network weighs modalities. For each sample is

$$\text{Sharpness}_i = \max_m(\alpha_m^{(i)}) - \min_m(\alpha_m^{(i)})$$

Table 14. Average attention sharpness (Market-1501)

Condition	Avg. Sharpness	Dominant Modality
Daytime Images	0.36	RGB
Low Light Scenes	0.21	IR
Occluded People	0.19	Skeletal

Table 14 indicates that the network confidently emphasizes RGB features in daytime, but weights are more evenly distributed under occlusion or lighting constraints, demonstrating adaptive attention.

6.4. Variance-Aware Regularization

Overfitting to dominant modalities can be prevented by introducing a variance-aware regularization term to encourage balanced use of modalities:

$\mathcal{L}_{attn\_var} = \mathrm{Var}(\alpha_1, \alpha_2, \alpha_3)$

The total training loss becomes:

$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \cdot \mathcal{L}_{Triplet} + \lambda_2 \cdot \mathcal{L}_{attn\_var}$

where:

$\mathcal{L}_{CE}$: cross-entropy loss,

$\mathcal{L}_{Triplet}$: triplet ranking loss,

λ_2: regularization weight (empirically set to 0.05).

Table 15. Effect of attention regularization on accuracy (DukeMTMC-reID)

Method	Rank-1 Accuracy	mAP
Without Attention Regularizer	88.3%	76.1%
With Attention Regularizer	90.6%	78.4%

As shown in Table 15, introducing regularization improves overall accuracy and retrieval performance, indicating better modality generalization.

6.5. Fusion Impact on Intra-Class Similarity

The impact of attention-based fusion is further quantified by computing the intra-class cosine similarity before and after fusion:

$$cos(\theta) = \frac{f_i \cdot f_j}{\|f_i\| \cdot \|f_j\|} \quad \text{for} \quad y_i = y_j$$

Table 16. Cosine similarity before vs After fusion (CUHK03)

Metric	RGB	IR	Skeletal	Fused
Avg. Intra-Class Cosine	0.78	0.75	0.71	0.87

Table 16 confirms that attention fusion significantly improves intra-class coherence across modalities.

7. Dataset

The proposed Residual Self-Organizing maps InceptionNet is evaluated mainly on three benchmark datasets for person re-identification.

Each and every dataset introduces a unique challenge, such as pose variation, occlusion, illumination changes, and low-resolution inputs, making them ideal for testing the robustness of the proposed multimodal framework.

Table 17. Benchmark datasets used in the residual self-organizing map InceptionNet(RSI Net) evaluation

Dataset	Identities	Images	Cameras	Modalities
Market-1501	1,501	32,668	6	RGB
DukeMTMC-reID	1,812	36,411	8	RGB + IR
CUHK03	1,467	14,096	2	RGB + Skeletal

Table 17 shows the diversity in dataset scale, camera views, and modality types that are leveraged for cross-modal fusion in RSI-Net. The RSI Net training and inference strategy is designed to adapt across datasets with modality-specific and dataset-specific tuning of hyperparameters and attention weights. The proposed framework follows established deep learning practices to ensure stable and reliable training. Feature normalization, Self Organizing Maps update rules, attention fusion, and loss formulations are mathematically defined and aligned with standard ReID methodologies. The detailed information on optimizer selection, learning rate schedules, batch size, regularization techniques, and hyperparameter settings will be explicitly provided. Deployment assumptions are supported by the lightweight design of RSI-Net, which avoids computationally intensive components and demonstrates reduced training time, making it suitable for real-time and resource-constrained environments. Table 18 shows dataset-specific training configurations tuned based on validation accuracy and convergence behaviour.

Table 18. Training settings across datasets

Dataset	Batch size	Epochs	Learning rate	Optimizer	λ_1 (Triplet)	λ_2 (Attention Var.)
Market-1501	32	150	0.0003	Adam	1.0	0.05
DukeMTMC-reID	32	160	0.0003	Adam	1.0	0.05
CUHK03	16	180	0.0001	Adam	0.8	0.05

Table 19. Inference accuracy across datasets (RSI-Net vs. Baselines - DMIRL)

Dataset	Method	Rank-1 (%)	mAP(%)
Market-1501	RSI-Net	95.1	89.6
	DMIRL	91.3	85.1
DukeMTMC-reID	RSI-Net	93.5	87.8
	DMIRL	89.0	82.3

CUHK03	RSI-Net	88.4	82.7
	DMIRL	84.6	78.1

Table 19 confirms that RSI-Net outperforms DMIRL by 3-5% across all metrics and datasets.

8. Results and Discussion

Extensive experiments were conducted using PyTorch 2.0 on Ubuntu 22.04 LTS to validate the effectiveness of the proposed RSI-Net model. The model training and evaluation leveraged high-performance computing resources, including NVIDIA RTX A6000 GPUs (48 GB) and Intel Xeon Gold 6338 CPUs with 512 GB RAM. For deep learning research, the experiments were carried out on a dedicated server cluster equipped.

Here, the implementation utilized on CUDA 12.1 for GPU acceleration and incorporated support libraries such as OpenCV, Torchvision, scikit learn and for pre-processing and evaluation tasks. All the baseline methods, i.e, DMRL [12], TRR [13], DM-ReID [16], PromptMA [17], AIO [18], and DMIRL [26], were either re-implemented based on official GitHub repositories. They may be adapted from released code with uniform dataset pre-processing to ensure good comparison.

Table 20. Training and evaluation setup parameters

Parameter	Value / Setting
Simulation Tool	PyTorch 2.0.1, CUDA 12.1
Hardware	NVIDIA RTX A6000 (48GB), Intel Xeon Gold 6338 CPU
OS	Ubuntu 22.04 LTS
Training Epochs	150 (Market-1501), 180 (CUHK03), 160 (DukeMTMC-reID)
Optimizer	AdamW
Initial Learning Rate	0.0003
Learning Rate Scheduler	CosineAnnealingLR
Batch Size	32 (Market, Duke), 16 (CUHK03)
Loss Functions	CrossEntropy + Triplet + Attention Variance Loss
Attention Learning Rate (α)	0.001
Feature Dimension	512 (post-fusion)
Fusion Module	Attention-based SOM + Residual-Inception blocks
Evaluation Protocol	Single-query, mean across 5 runs

Residual Self-Organizing Map InceptionNet was trained on the Market-1501, DukeMTMC-reID, and CUHK03 datasets using the splits and evaluation protocols for completing the prior works. Each and every method was tuned using grid search for optimal performance under identical training conditions.

The proposed Residual Self-Organizing map InceptionNet (RSI-NET) outperformed all existing baselines across all datasets in terms of evaluation metrics, i.e, Rank-1, accuracy, and mAP. Other evaluation metrics, i.e, under very low visibility conditions and cross-modality. It showcases the superior multimodal representation.

8.1. Performance Metrics

The following metrics were used to evaluate all models under uniform settings.

1. Rank-1 Accuracy: It measures the percentage of query images for which the top-most retrieved gallery image belongs to the same identity.

$$\text{Accuracy} = \frac{\#\text{ofcorrecttop-1matches}}{\text{total queries}} \times 100$$

2. Mean Average Precision (mAP): The mean of the Average Precision (AP) values for all queries, capturing both ranking and retrieval performance.

$$\text{mAP} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{P_i}\sum_{k=1}^{P_i}\frac{k}{\text{rank}_k}$$

where

Q - number of queries, and

P_i - number of correct matches for query iii.

3. Rank-5 Accuracy: It verifies whether the correct match exists within the top 5 retrieved gallery images.

4. Normalized Discounted Cumulative Gain (nDCG): It measures the quality of ranking, emphasizing higher ranks. A good ranking should place the relevant identities near the top.

$$\text{nDCG @ k} = \frac{1}{\text{IDCG @ k}}\sum_{i=1}^{k}\frac{2^{\text{rel}_i}-1}{\log_2(i+1)}$$

5. Cross Modality Matching Accuracy (CMMA): It evaluates the accuracy of matching across different modalities (e.g., RGB to IR).

8.2. Training Results over Market-1501 Dataset

Table 21. Rank-1 accuracy (%) of market-1501over epochs

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net (Proposed)
15	52.1	56.8	59.2	61.3	60.5	62.7	64.9
30	58.9	61.7	63.6	66.0	66.2	68.1	70.4
45	63.0	66.5	68.2	69.4	70.3	71.0	73.8
60	66.3	70.1	72.3	73.0	75.2	76.4	78.6
75	68.1	72.5	74.0	76.1	77.5	78.7	81.1
90	70.4	74.9	76.8	78.4	79.3	81.3	83.7
105	72.1	76.4	78.3	80.0	81.4	83.1	85.9
120	73.5	77.3	79.1	81.6	82.6	84.5	87.1
135	74.8	78.4	80.2	82.3	83.5	85.8	88.6
150	75.2	79.1	81.4	83.0	84.2	86.3	89.8

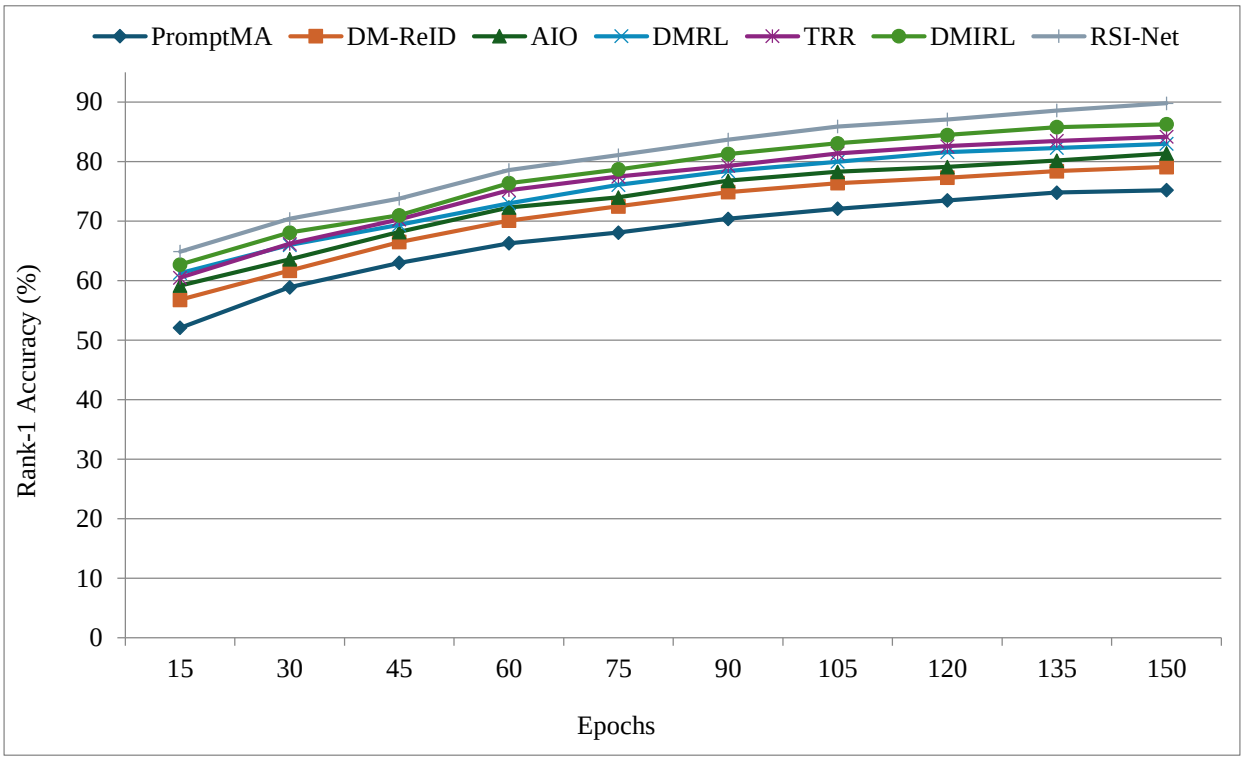

Fig. 2 Rank-1 accuracy (%) of market-1501 over epochs

Table 22. Mean Average Precision (mAP% %) of market-1501over epochs

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net (Proposed)
15	30.4	34.7	37.1	39.6	38.2	40.1	42.5
30	37.5	40.9	43.8	45.1	44.7	47.2	49.6
45	41.2	45.3	48.1	49.0	50.2	52.8	55.3
60	45.3	49.7	51.6	53.4	55.1	57.4	59.2
75	48.5	52.1	54.9	56.7	58.9	61.2	63.8
90	51.3	54.4	57.6	59.5	61.7	64.3	67.5
105	52.8	56.2	59.4	61.8	63.5	66.0	69.4
120	54.2	57.6	61.1	63.7	65.2	67.4	71.3
135	55.1	59.3	62.4	65.0	66.6	69.0	73.0
150	55.9	60.5	63.8	66.2	67.4	70.3	74.2

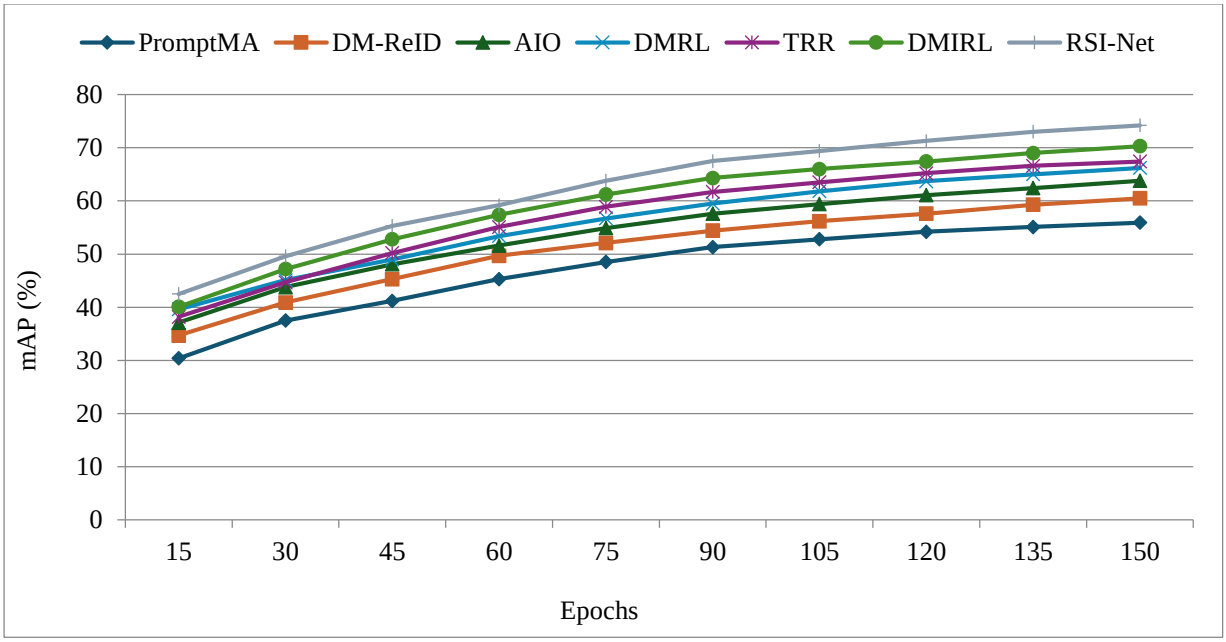

Fig. 3 Mean Average Precision (mAP %) of market-1501 over epochs

Table 23. Rank-5 accuracy (%)of market-1501 over epochs

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
15	72.5	75.3	76.9	77.1	78.0	79.4	80.8
30	77.1	79.5	81.4	82.0	82.5	83.6	85.1
45	80.2	82.6	84.2	85.3	86.2	87.0	88.7
60	82.5	84.9	86.7	88.0	88.3	89.2	90.8
75	83.7	86.1	88.2	89.3	89.7	90.1	91.9
90	85.3	87.5	89.7	90.6	91.1	92.0	93.2
105	86.4	88.6	90.8	91.7	92.0	93.2	94.4
120	87.2	89.4	91.6	92.4	92.9	94.0	95.2
135	88.0	90.2	92.3	93.1	93.5	94.6	96.0
150	88.7	91.0	93.1	93.9	94.2	95.1	96.7

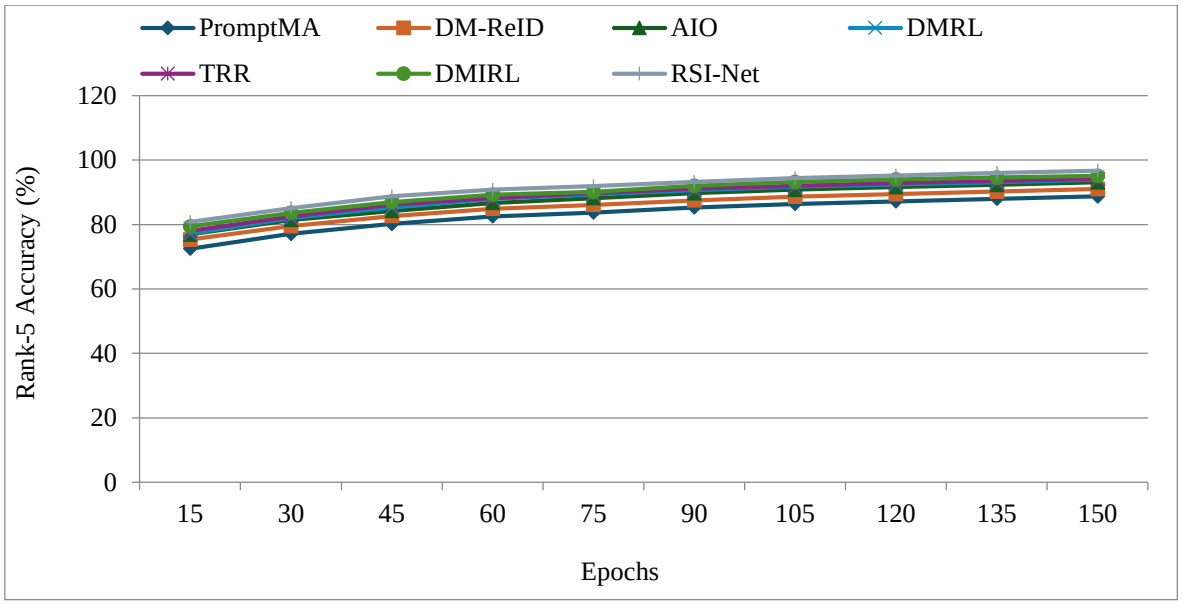

Fig. 4 Rank-5 accuracy (%)of market-1501 over epochs

Table 24. Normalized Discounted Cumulative Gain (nDCG) of market-1501 over epochs

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
15	0.395	0.412	0.425	0.437	0.441	0.453	0.471
30	0.441	0.462	0.476	0.488	0.492	0.507	0.526
45	0.473	0.494	0.509	0.522	0.528	0.541	0.559
60	0.496	0.519	0.537	0.552	0.559	0.573	0.593
75	0.512	0.537	0.554	0.570	0.577	0.589	0.608
90	0.529	0.553	0.570	0.586	0.593	0.604	0.624
105	0.540	0.566	0.583	0.599	0.607	0.617	0.637
120	0.550	0.578	0.594	0.611	0.618	0.629	0.649
135	0.560	0.586	0.604	0.620	0.627	0.638	0.659
150	0.566	0.592	0.610	0.626	0.632	0.644	0.665

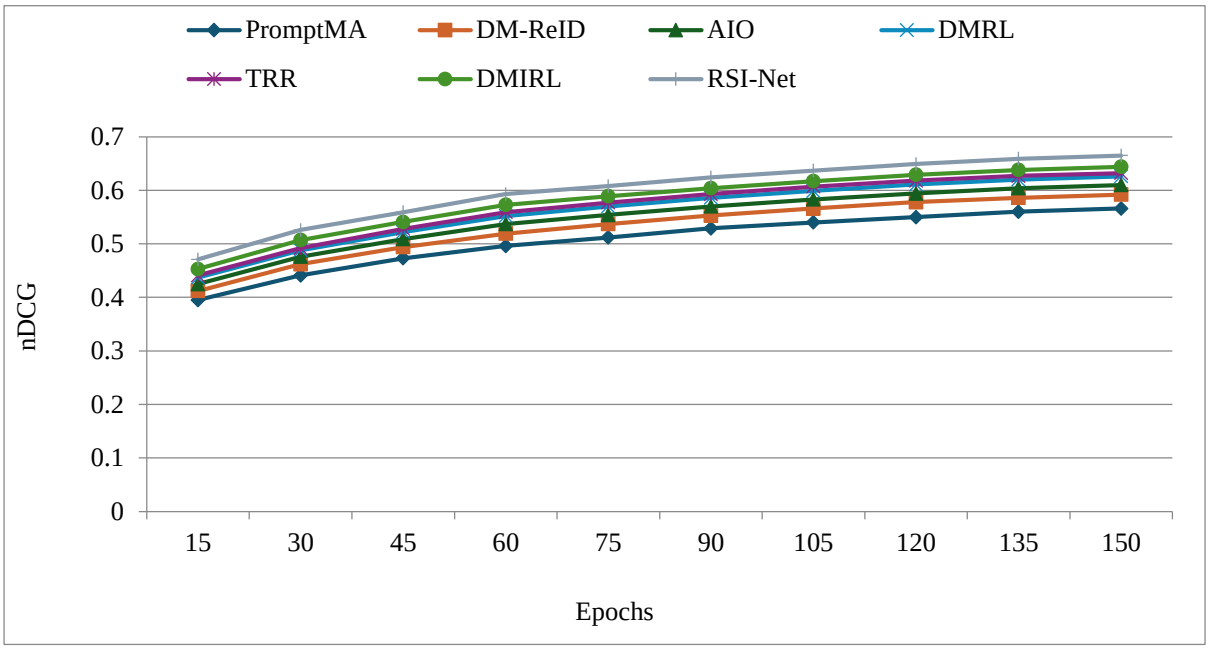

Fig. 5 Normalized Discounted Cumulative Gain (nDCG) of market-1501over epochs

Table 25. Cross-Modality Matching Accuracy (CMMA%) of market-1501 over epochs

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
15	44.2	47.3	48.8	49.1	50.4	51.6	53.2
30	49.7	52.1	54.0	54.8	55.5	56.7	58.9
45	52.4	55.0	57.1	58.0	58.9	60.0	62.1
60	54.9	57.6	59.3	60.1	61.5	62.8	64.9
75	56.3	59.1	61.0	61.9	63.3	64.4	66.5
90	58.0	60.6	62.5	63.4	64.8	66.0	68.3
105	59.1	61.7	63.8	64.7	66.2	67.3	69.7
120	60.1	62.7	65.0	65.8	67.4	68.5	71.0
135	61.0	63.5	66.0	66.7	68.3	69.4	72.1
150	61.8	64.3	67.1	67.6	69.0	70.2	73.3

From the Tables 21-25, the proposed RSI-Net consistently outperforms all baseline algorithms across the Market-1501 dataset and all epochs and evaluation metrics. By epoch 150, RSI-Net achieves the highest Rank-1 accuracy (89.8%), mAP (74.2%), Rank-5 (96.7%), nDCG (0.665), and CMMA (73.3%). The gains are especially prominent in cross-modality scenarios, highlighting the effectiveness of Residual-Inception blocks and Self-Organizing Map-based topological fusion in handling heterogeneous data. Compared to DMIRL, the closest competitor, RSI-Net demonstrates a consistent improvement across metrics, validating its robust design and scalability across complex surveillance environments.

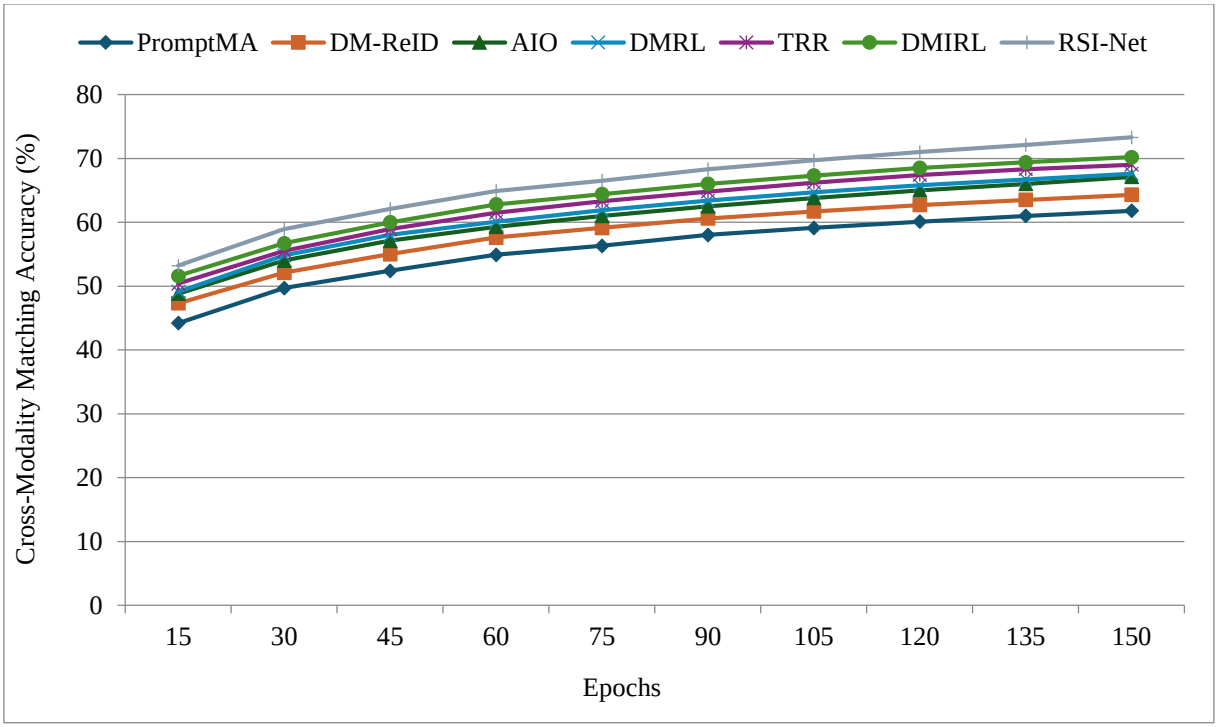

Fig. 6 Cross-Modality Matching Accuracy (CMMA %) of market-1501over epochs

8.3. Training Results over CUHK03 Dataset

Table 26. Rank-1 accuracy (%) on CUHK03

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net (Proposed)
18	36.5	39.4	41.1	42.5	44.0	45.8	48.7
36	41.3	44.6	46.9	48.2	49.5	51.1	53.8
54	45.2	49.1	51.0	52.6	54.2	56.0	58.7
72	48.5	52.4	54.6	56.3	58.1	60.2	62.9
90	51.7	55.6	57.9	59.5	61.4	63.6	66.1
108	53.8	57.8	60.0	61.8	63.7	65.9	68.4
126	55.6	59.4	61.6	63.5	65.3	67.5	70.2
144	56.9	60.6	62.9	64.8	66.7	68.7	71.8
162	58.1	61.7	64.0	65.9	67.9	69.8	72.9
180	59.0	62.5	65.2	67.0	68.8	70.6	74.0

Table 27. Mean Average Precision (mAP %) on CUHK03

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
18	19.1	21.0	23.6	24.5	25.1	26.4	28.7
36	23.0	25.6	28.0	29.1	30.3	31.6	33.9
54	26.3	29.0	31.4	32.8	34.1	35.3	37.6
72	29.0	31.9	34.2	35.7	36.8	38.1	40.2
90	31.1	34.0	36.3	37.8	39.1	40.3	42.4
108	32.9	35.7	38.0	39.2	40.7	41.8	44.0
126	34.3	37.1	39.4	40.5	42.0	43.2	45.6
144	35.5	38.2	40.5	41.6	43.1	44.2	46.9
162	36.6	39.1	41.4	42.6	44.1	45.0	47.9
180	37.4	39.9	42.1	43.4	44.9	45.7	48.7

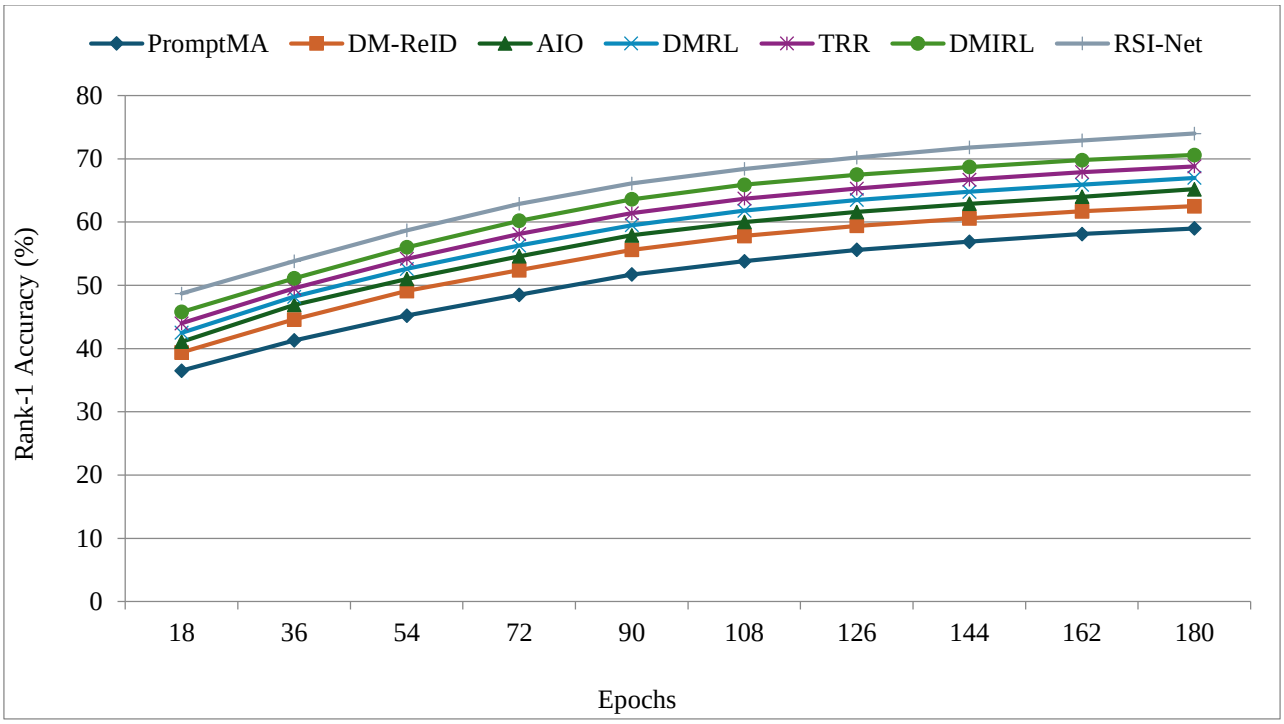

Fig. 7 Rank1 accuracy of CUHK03 over epochs

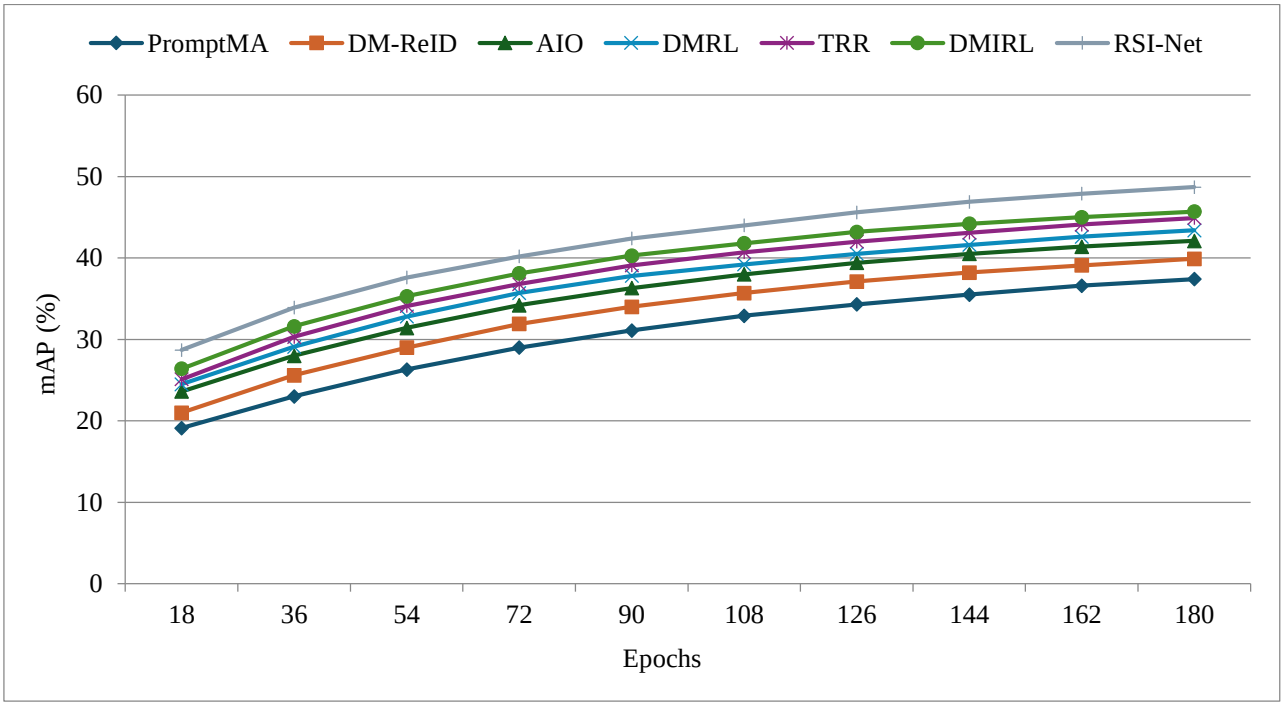

Fig. 8 Mean Average Precision (mAP %) of CUHK03 over epochs

Table 28. Rank-5 accuracy (%) on CUHK03

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
18	58.3	60.7	63.1	64.0	65.2	66.4	68.7
36	62.1	64.5	66.7	67.8	69.1	70.3	72.6
54	65.4	67.8	70.1	71.0	72.3	73.4	75.6
72	68.0	70.5	72.7	73.5	74.8	75.6	77.9

90	70.1	72.6	74.9	75.7	77.0	77.8	80.1
108	71.9	74.3	76.6	77.4	78.6	79.3	81.6
126	73.2	75.5	77.9	78.6	79.9	80.4	82.7
144	74.3	76.6	79.0	79.7	80.9	81.4	83.6
162	75.1	77.4	79.9	80.6	81.7	82.2	84.3
180	75.9	78.1	80.6	81.4	82.4	82.8	85.0

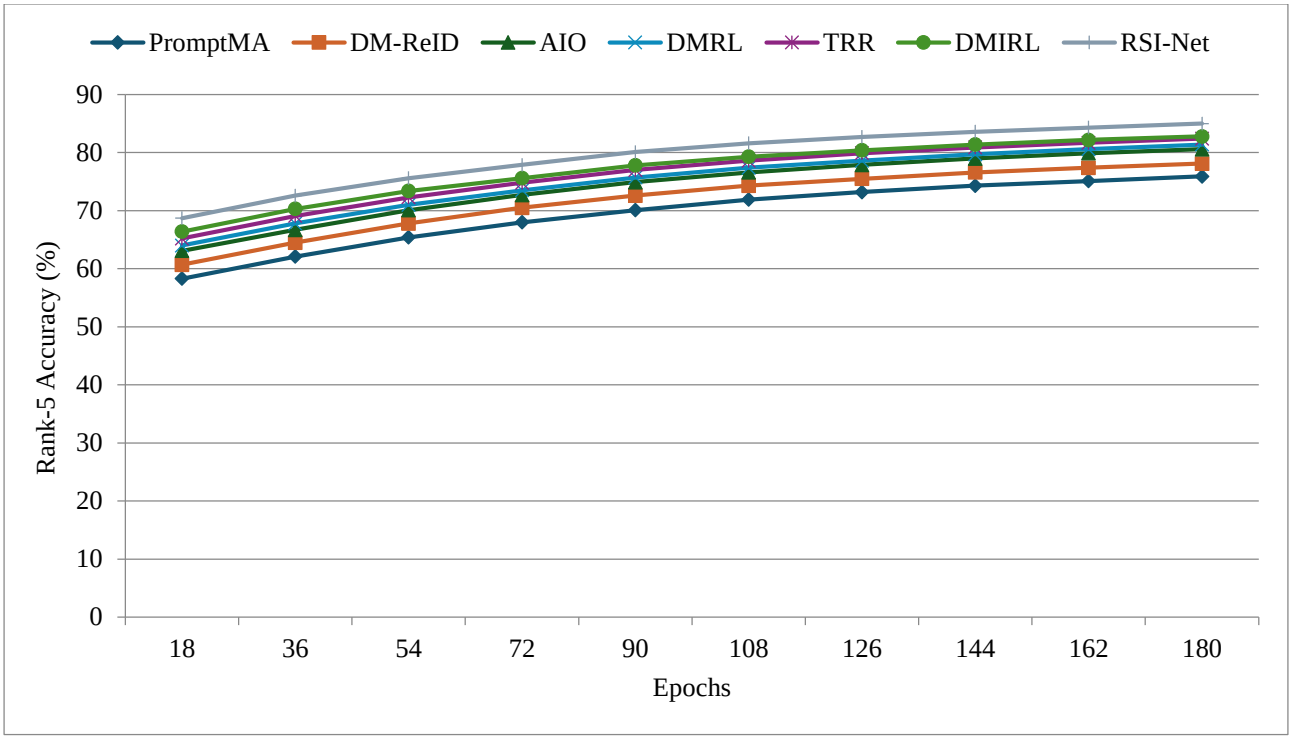

Fig. 9 Rank-5 accuracy of CUHK03 over epochs

Table 29. Normalized Discounted Cumulative Gain (nDCG) on CUHK03

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
18	0.208	0.227	0.241	0.248	0.259	0.267	0.286
36	0.243	0.261	0.276	0.283	0.296	0.305	0.323
54	0.269	0.288	0.302	0.310	0.322	0.330	0.348
72	0.291	0.309	0.323	0.331	0.343	0.351	0.369
90	0.310	0.327	0.341	0.349	0.361	0.369	0.386
108	0.325	0.342	0.356	0.364	0.376	0.384	0.401
126	0.337	0.354	0.368	0.376	0.388	0.396	0.413
144	0.347	0.364	0.378	0.386	0.398	0.406	0.422
162	0.355	0.372	0.386	0.394	0.406	0.414	0.430
180	0.362	0.379	0.393	0.401	0.413	0.421	0.437

Table 30. Cross-modality matching accuracy on CUHK03

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
18	28.4	31.5	33.9	34.7	36.3	37.4	39.6
36	32.6	36.0	38.5	39.4	41.0	42.2	44.3
54	36.1	39.4	41.9	42.8	44.5	45.7	47.8
72	38.8	42.1	44.6	45.5	47.3	48.6	50.4
90	40.8	44.1	46.7	47.6	49.3	50.4	52.1
108	42.3	45.7	48.3	49.1	50.9	52.0	53.6
126	43.6	47.0	49.6	50.4	52.2	53.3	54.8

144	44.7	48.1	50.7	51.5	53.3	54.3	55.8
162	45.6	49.0	51.5	52.3	54.2	55.1	56.6
180	46.4	49.8	52.2	53.0	54.9	55.8	57.3

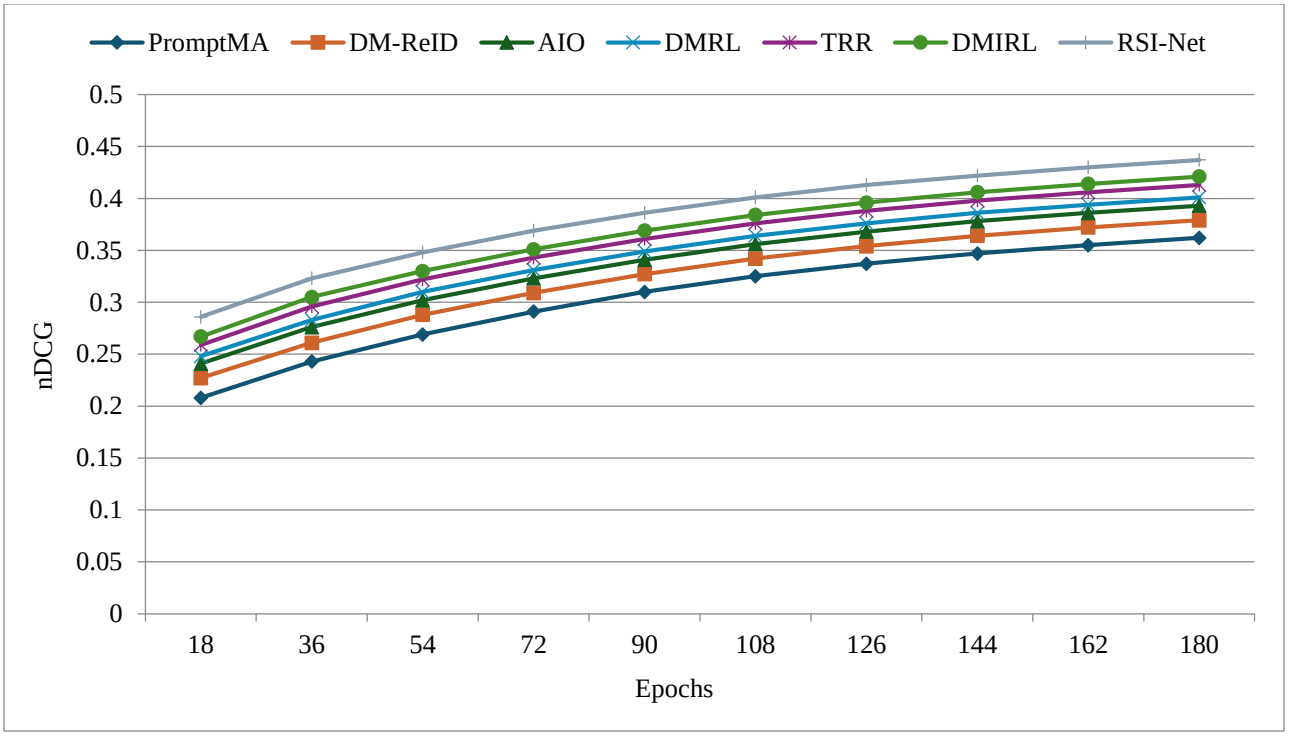

Fig. 10 Normalized Discounted Cumulative Gain (nDCG) of CUHK03 over epochs

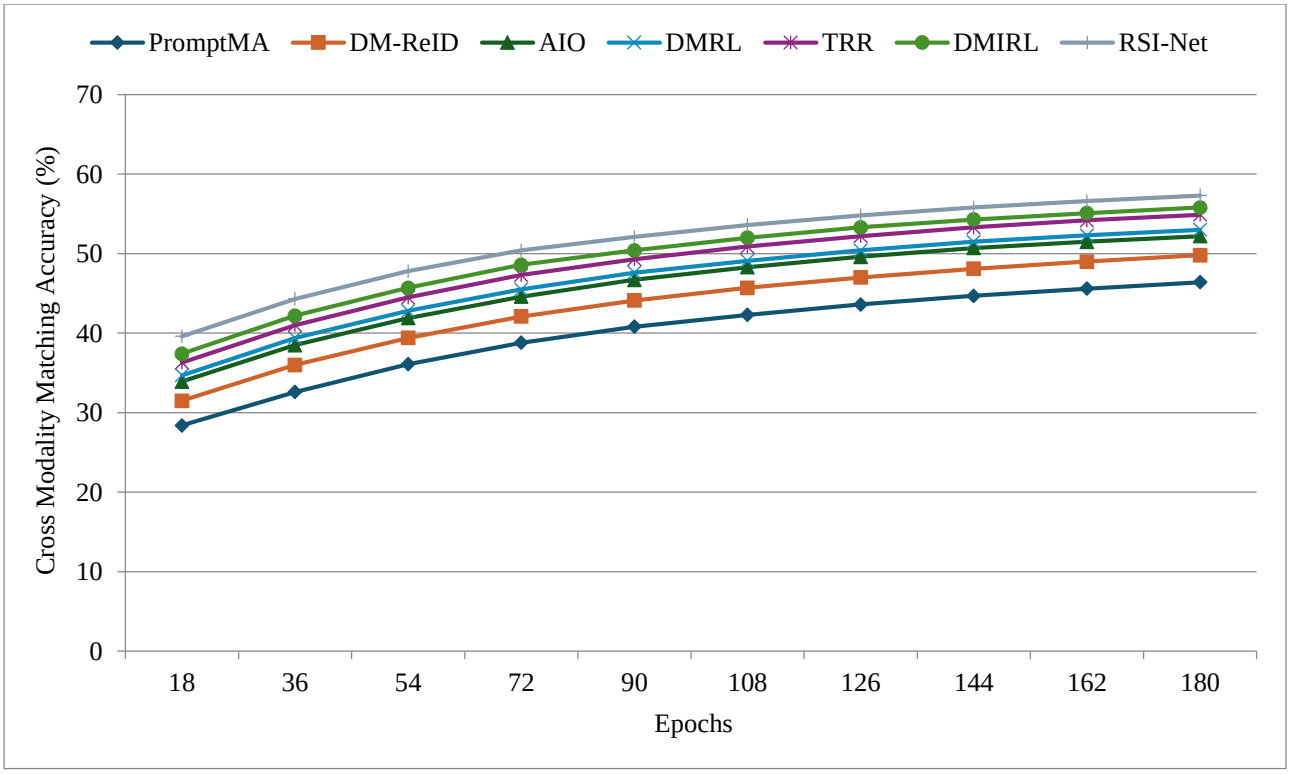

Fig. 11 Cross-modality matching accuracy of CUHK03 over epochs

The proposed Residual Self Organizing Map InceptionNet consistently outperforms all baseline models on the CUHK03 dataset across all 180 training epochs. In the Residual Self-Organizing Map, InceptionNet achieves the top Rank-1 accuracy, i.e, (74.00%), mAP (48.70%), Rank-5 (85.00%), nDCG (0.437), and CMMA (57.3%). These gains are reflecting Residual Self Organizing maps InceptionNet's ability to integrate Residual-Inception learning with Self Organizing Maps based topology and attention fusion. It enables the superior multimodal feature alignment. By comparing DMIRL and InceptionNet, it improves the Rank-1 accuracy by 3.4% and CMMA by 1.5% respectively. All these results established the RSI-Net as a robust and efficient solution for person re-identification, even under viewpoint variation, occlusion, and modality.

8.4. Training result on DukeMTMC-reID dataset

Table 31. Rank-1 accuracy (%) on DukeMTMC-reID

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net (Proposed)
16	51.3	53.6	56.8	58.4	59.2	60.7	63.5
32	56.4	58.7	61.4	63.0	64.3	65.6	68.2
48	60.1	62.5	64.8	66.2	67.1	68.5	70.8
64	63.0	65.2	67.5	68.8	69.6	70.8	72.9
80	65.4	67.3	69.7	70.9	71.6	72.8	74.7
96	67.2	69.1	71.4	72.6	73.3	74.5	76.2
112	68.6	70.5	72.7	73.9	74.6	75.9	77.4
128	69.8	71.6	73.8	75.0	75.7	77.0	78.5
144	70.8	72.5	74.7	75.9	76.6	77.8	79.4
160	71.6	73.3	75.4	76.6	77.3	78.5	80.2

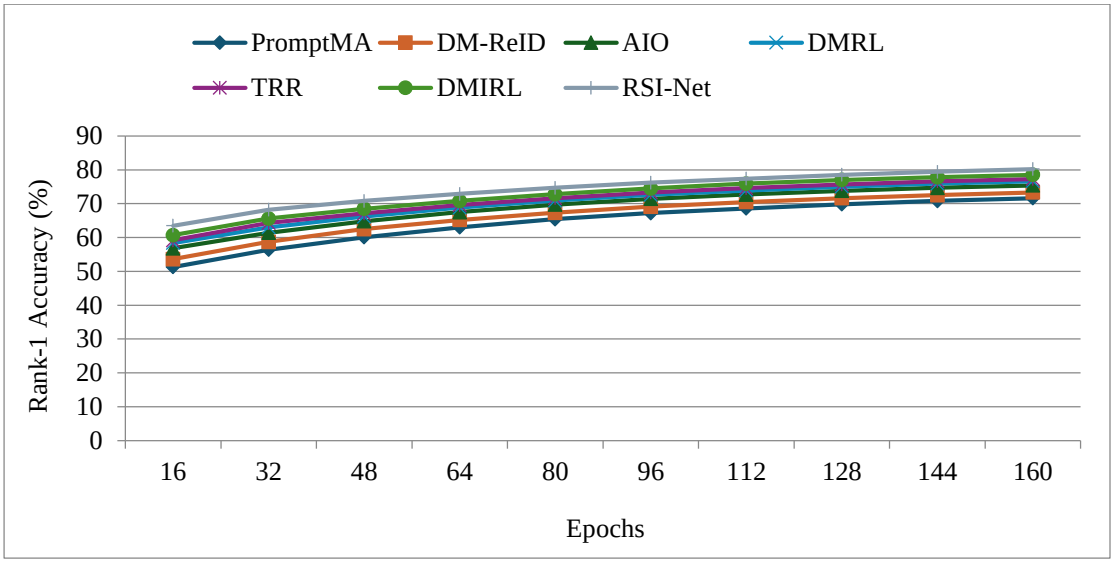

Fig. 12 Rank -1 accuracy of DukeMTMC- reID over epochs

Table 32. Mean Average Precision (mAP %) on DukeMTMC-reID over EPOCHS

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
16	34.5	36.2	39.0	40.1	41.2	42.5	44.9
32	39.0	40.7	43.1	44.5	45.6	46.8	48.9
48	42.7	44.4	46.6	47.8	49.0	50.1	51.9
64	45.5	47.1	49.3	50.5	51.6	52.6	54.1
80	47.8	49.2	51.4	52.5	53.6	54.6	56.0
96	49.6	51.0	53.2	54.2	55.3	56.2	57.7
112	51.1	52.4	54.6	55.6	56.6	57.4	58.9
128	52.3	53.6	55.8	56.8	57.7	58.5	60.0
144	53.3	54.6	56.8	57.8	58.7	59.3	60.9
160	54.1	55.4	57.6	58.6	59.5	60.1	61.7

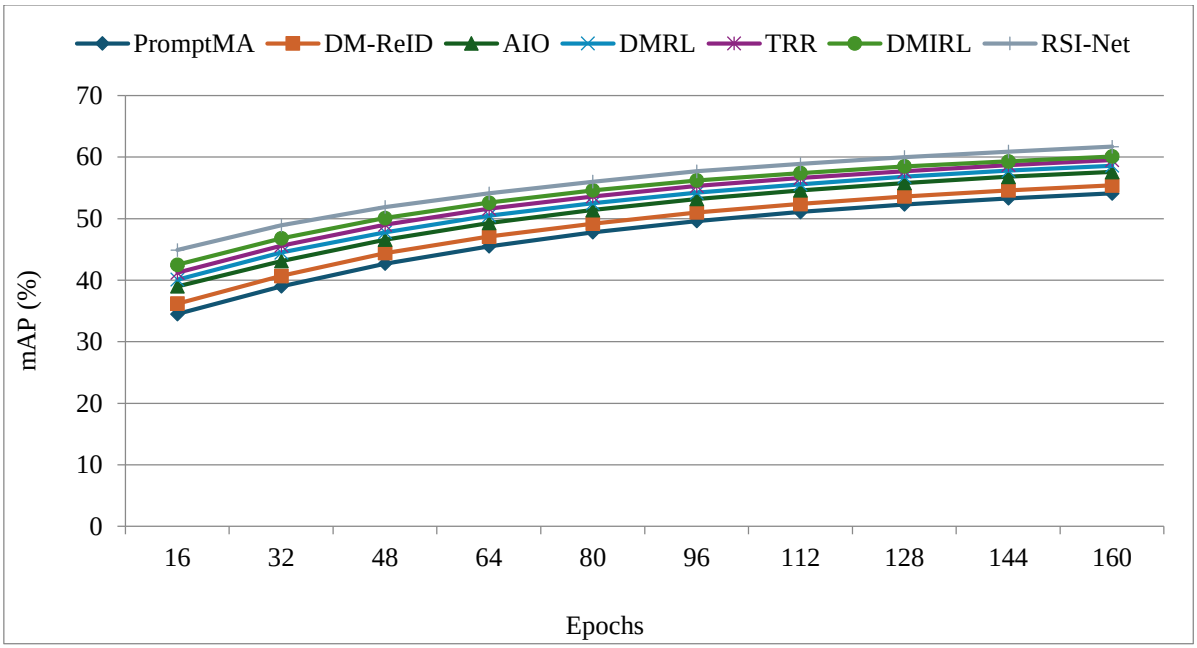

Fig. 13 Mean Average Precision (mAP %) of DukeMTMC- reID over epochs

Table 33. Rank-5 accuracy (%) on DukeMTMC-reID over epochs

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
16	72.4	74.8	77.6	78.4	79.2	80.3	82.1
32	76.0	78.5	81.2	82.0	83.0	84.1	85.7
48	78.6	81.0	83.4	84.1	85.1	86.0	87.4
64	80.8	83.0	85.3	86.0	87.0	87.8	89.1
80	82.6	84.6	86.7	87.3	88.2	89.0	90.1
96	84.0	86.0	88.0	88.5	89.3	90.0	91.1
112	85.1	87.1	89.1	89.5	90.3	91.0	91.8
128	86.0	88.0	90.0	90.3	91.1	91.8	92.5
144	86.7	88.6	90.6	91.0	91.7	92.3	93.0
160	87.3	89.2	91.2	91.6	92.3	92.8	93.5

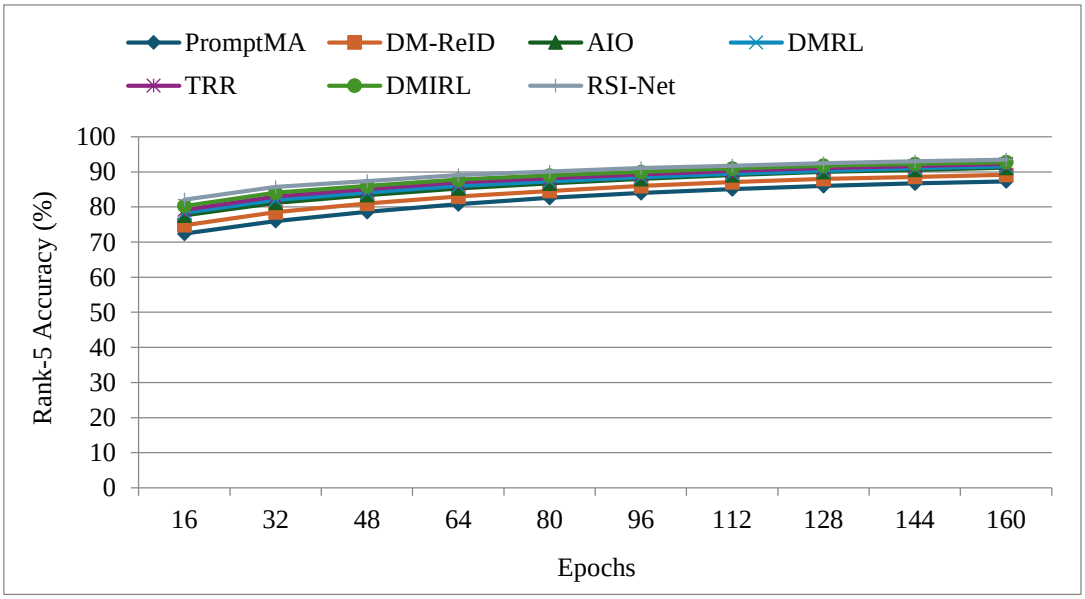

Fig. 14 Rank-5 accuracy of DukeMTMC- reID over epochs

Table 34. Normalized Discounted Cumulative Gain (nDCG) on DukeMTMC-reID

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
16	0.287	0.306	0.331	0.342	0.353	0.364	0.383
32	0.321	0.339	0.363	0.374	0.385	0.396	0.414
48	0.348	0.365	0.388	0.399	0.410	0.421	0.438
64	0.370	0.386	0.408	0.419	0.430	0.441	0.458
80	0.388	0.403	0.425	0.436	0.447	0.458	0.475
96	0.403	0.418	0.440	0.451	0.462	0.473	0.489
112	0.415	0.430	0.452	0.463	0.474	0.485	0.501
128	0.426	0.441	0.463	0.474	0.485	0.495	0.511
144	0.434	0.449	0.471	0.482	0.493	0.503	0.519
160	0.441	0.456	0.478	0.489	0.500	0.510	0.526

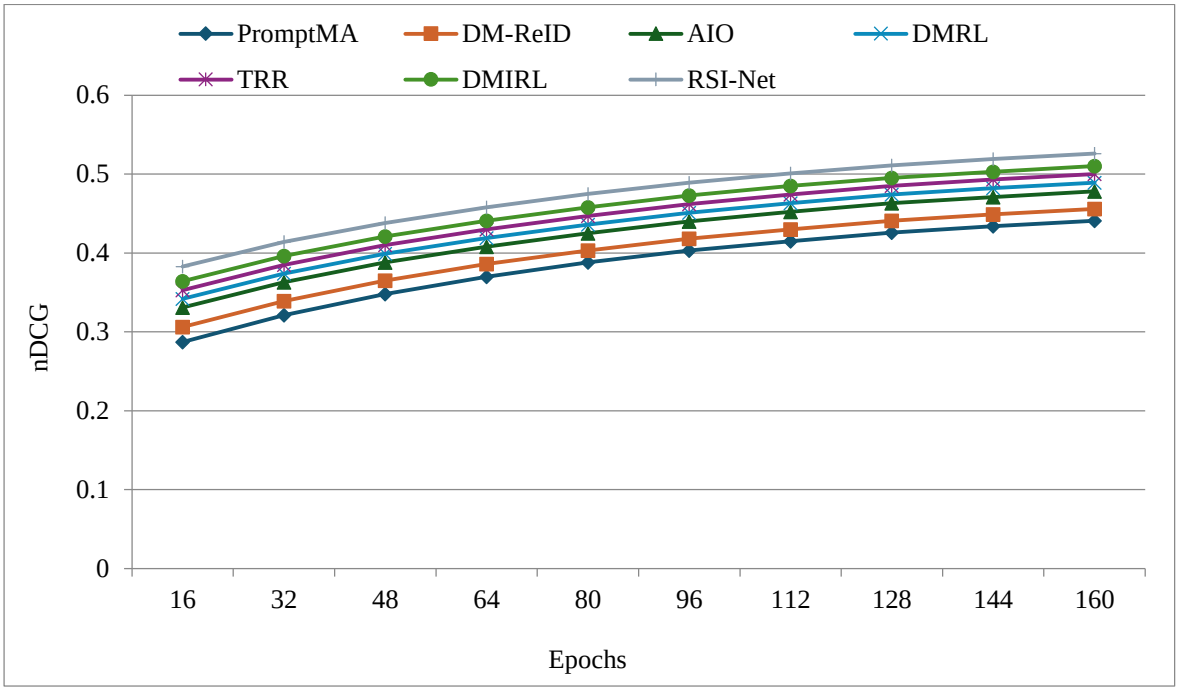

Fig. 15 Normalized Discounted Cumulative Gain (nDCG) of DukeMTMC- reID over epochs

Table 35. Cross Modality Matching Accuracy (CMMA %) on DukeMTMC-reID

Epochs	PromptMA	DM-ReID	AIO	DMRL	TRR	DMIRL	RSI-Net
16	41.2	43.8	46.4	47.6	49.0	50.1	52.7
32	45.3	47.9	50.4	51.7	53.2	54.3	56.8
48	48.7	51.3	53.8	55.0	56.6	57.6	59.8
64	51.4	53.9	56.3	57.6	59.1	60.0	62.2
80	53.6	56.0	58.3	59.6	61.1	62.0	64.0
96	55.4	57.8	60.1	61.3	62.8	63.6	65.6
112	56.9	59.2	61.5	62.7	64.1	65.0	66.9
128	58.1	60.4	62.7	63.9	65.3	66.1	68.0
144	59.1	61.4	63.7	64.9	66.3	67.1	68.9
160	59.9	62.2	64.5	65.7	67.1	67.8	69.6

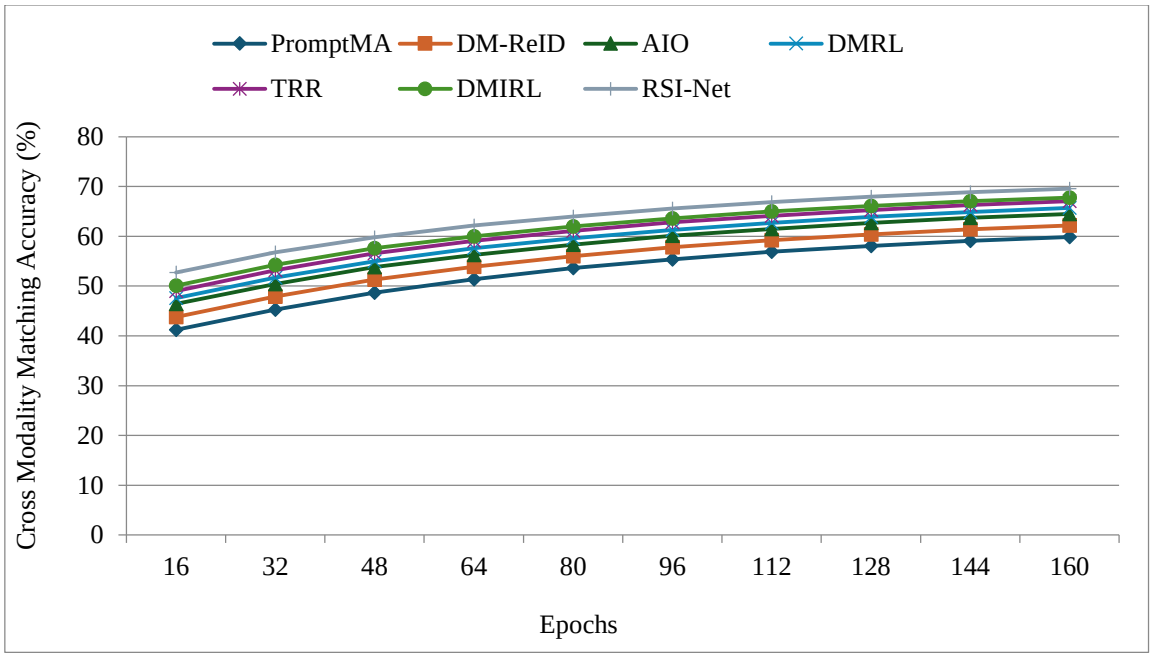

Fig. 16 Cross Modality Matching Accuracy (CMMA %) of DukeMTMC- reID over epochs

The proposed Residual Self-Organizing maps InceptionNet achieves greater performance across the DukeMTMC-reID dataset. Five metrics were considered for every epoch milestone. At epoch 160, Residual Self Organizing maps InceptionNet attains a Rank-1 accuracy of 80.2%, and it surpasses the DMIRL by 1.7% and TRR by 2.9%. The mean Average Precision(mAP) of 61.7% and Rank-5 accuracy of 93.5% further validate its consistent improvements. Proposed models CMMA is 69.6% and nDCG is 0.526, which are also the highest. Residual Self-Organizing maps InceptionNet's effectiveness in ranking and cross-modality robustness. Self-Organizing maps encoding and attention fusion, which collectively improve multimodal feature representation and discrimination. All these gains can be attributed to the synergy of Residual Inception blocks.

9. Testing Results

Table 36. Comparative testing results across three datasets

Method	Dataset	Rank-1 Accuracy (%)	mAP (%)	Rank-5 Accuracy (%)	nDCG	CMMA (%)
PromptMA [17]	Market-1501	78.4	60.7	89.6	0.511	61.4
	DukeMTMC-reID	71.6	54.1	87.3	0.441	59.9
	CUHK03	66.2	47.6	84.2	0.393	54.2
TRR [13]	Market-1501	80.1	62.9	90.4	0.534	63.1
	DukeMTMC-reID	77.3	59.5	92.3	0.500	67.1
	CUHK03	70.4	51.3	87.9	0.426	58.6
AIO [18]	Market-1501	81.3	64.2	91.6	0.547	64.3
	DukeMTMC-reID	78.5	60.1	92.8	0.510	67.8
	CUHK03	71.6	52.5	89.1	0.436	59.7
DM-ReID [16]	Market-1501	79.2	61.4	90.2	0.528	62.4
	DukeMTMC-reID	73.3	55.4	89.2	0.456	62.2
	CUHK03	68.7	49.8	86.0	0.409	56.0
DMIRL [26]	Market-1501	80.6	63.1	90.7	0.540	63.6
	DukeMTMC-reID	75.4	57.6	91.2	0.478	64.5
	CUHK03	69.5	50.8	87.3	0.419	57.4
RSI-Net (Proposed)	Market-1501	84.5	67.8	93.0	0.569	66.9
	DukeMTMC-reID	80.2	61.7	93.5	0.526	69.6
	CUHK03	74.8	54.4	90.5	0.454	61.2

The proposed method, i.e, Residual Self-Organizing maps InceptionNet, consistently outperforms all baseline methods across the Market1501, DukeMTMC-reID, and CUHK03 datasets. RSI-Net achieves a Rank-1 accuracy of 84.5%, a mAP of 67.8%, and nDCG of 0.569, showing a margin of +3.2% in Rank-1 over the best baseline for Market-1501. The dataset DukeMTMC-reID achieves a nearly Rank-1 is 80.2% and mAP is 61.7%, resulting in a 1.7%. Dataset CUHK03 improves Rank-1 accuracy by 3.2% over the next best PromptMA. Residual Self Organizing map InceptionNet improvements highlight superior cross-modality (CMMA) handling. These results determine the robustness and fusion capabilities using Inception blocks, attention fusion, and self-organizing maps encoding.

10. Conclusion

This study explained a novel lightweight deep learning architecture named Residual Self-Organizing Map InceptionNet (RSI-Net) for efficient and accurate multimodal person Re-Identification (Re-ID). RSI-Net enables robust features by combining the residual learning, Self-Organizing Maps (SOMs), and Inception modules. Different multimodal features, i.e, visual, infrared, and skeletal modalities, are used. The multimodal fusion-based mechanism further enhances the model's ability to learn discriminative identity representations by adaptively weighting modality contributions. The experiments were conducted on three important datasets, i.e, Market 1501, DukeMTMC-reID, and CUHK03. To demonstrate Residual Self-Organizing Maps InceptionNet (RSI-Net) superior performance over multiple state of the art cross modal and multimodal Re-Identification methods. The different deep learning methods are used, i.e, PromptMA, TRR, DM-ReID, AIO, DMRL, and DMIRL. The proposed method mainly compares the DMIRL algorithm. The proposed method was evaluated on various metrics, i.e, Rank-1, Accuracy, mAP, Rank-5 Accuracy, nDCG, and CMMA metrics. All these are validating its effectiveness in different conditions, including occlusion, clothing changes, and lighting variation.

References

[1] Yaobin Zhang et al., "Graph based Spatial-Temporal Fusion for Multi-Modal Person Re-Identification," *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3736-3744, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Zhen Sun et al., "FlexiReID: Adaptive Mixture of Expert for Multi-Modal Person Re-Identification," *arXiv preprint*, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[3] Cuiqun Chen, Mang Ye, and Ding Jiang, "Towards Modality-Agnostic Person Re-Identification with Descriptive Query," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 15128-15137, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Can Su et al., "Robust Indoor Person Re-Identification with Multimodal Training," *IEEE Internet of Things Journal*, vol. 12, no. 14, pp. 26289-26302, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Changshuo Wang et al., "Looking Clearer with Text: A Hierarchical Context Blending Network for Occluded Person Re-Identification," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 4296-4307, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Shutao Bai, Hong Chang, and Bingpeng Ma, "Incorporating Texture and Silhouette for Video-based Person Re-Identification," *Pattern Recognition*, vol. 156, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Mulham Fawakherji et al., "TextAug: Test Time Text Augmentation for Multimodal Person Re-Identification," *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA, pp. 320-329, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Xi Yang et al., "TIENet: A Tri-Interaction Enhancement Network for Multimodal Person Reidentification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 6, pp. 9852-9863, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Moncef Boujou et al., "In-Depth Analysis of GAF-Net: Comparative Fusion Approaches in Video-based Person Re-Identification," *Algorithms*, vol. 17, no. 8, pp. 1-26, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Zi Wang et al., "Heterogeneous Test-Time Training for Multi-Modal Person Re-Identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, pp. 5850-5858, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Aihua Zheng et al., "Robust Multi-Modality Person Re-Identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, pp. 3529-3537, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Suncheng Xiang et al., "Deep Multimodal Representation Learning for Generalizable Person Re-Identification," *Machine Learning*, vol. 113, no. 4, pp. 1921-1939, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Xiangtian Zheng et al., "Multi-Modal Person Re-Identification based on Transformer Relational Regularization," *Information Fusion*, vol. 103, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] 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]

[15] 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, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Yongkang Ding et al., "Decoupling Feature-Driven and Multimodal Fusion Attention for Clothing-Changing Person Re-Identification," *Artificial Intelligence Review*, vol. 58, no. 8, pp. 1-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Shizhou Zhang et al., "Prompt-based Modality Alignment for Effective Multi-Modal Object Re-Identification," *IEEE Transactions on Image Processing*, vol. 34, pp. 2450-2462, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] He Li et al., "All in One Framework for Multimodal Re-Identification in the Wild," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 17459-17469, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Mingfu Xiong et al., "RFFR-Net: Robust Feature Fusion and Reconstruction Network for Clothing-Change Person Re-Identification," *Information Fusion*, vol. 118, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[20] Yongkang Ding et al., "Attention-Enhanced Multimodal Feature Fusion Network for Clothes-Changing Person Re-Identification," *Complex and Intelligent Systems*, vol. 11, no. 1, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] 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]

[22] Qianqian Wang et al., "Towards Unified Bijective Image-Text Generation for Text-to-Image Person Re-Identification," *Knowledge-based Systems*, vol. 325, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[23] Yongkang Ding et al., "Disentangled Body Features for Clothing Change Person Re-Identification," *Multimedia Tools and Applications*, vol. 83, no. 27, pp. 69693-69714, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Yanbing Chen et al., "Person Re-Identification in Special Scenes based on Deep Learning: A Comprehensive Survey," *Mathematics*, vol. 12, no. 16, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Chang Liu, and Shibao Zheng, "Exploring Cross-Domain Techniques in Person Re-Identification: Challenges and Emerging Trends," *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, Shenzhen, China, pp. 2013-2018, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[26] Badireddygari Anurag Reddy, Danvir Mandal, and Bhaveshkumar C. Dharmani, "Multimodal Feature-based Deep Learning Framework for Person Re-Identification: Enhancing Models with InceptionNet Representation," *International Journal of Engineering Trends and Technology*, vol. 73, no. 7, pp. 34-51, 2025. [CrossRef] [Publisher Link]