Enhancing Target Re-Identification via Model Fusion and Knowledge Distillation of Pre-trained Foundation Models

Tendai Shoko[1[0000−0002−1245−4267]] and Terence L. van Zyl[1,2[0000−0003−4281−630X]]

[1] Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa
223029866@student.uj.ac.za
[2] CAIR, Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa
tvanzyl@uj.ac.za

Abstract. Target re-identification (re-ID) systems face critical deployment challenges balancing accuracy with computational efficiency in resource-constrained environments. This paper presents a novel framework integrating Mixture-of-Experts (MoE) with Knowledge Distillation (KD) to leverage pre-trained foundation models effectively. The framework employs dynamic expert selection to combine CLIP and ALIGN models, then distills their collective knowledge into a compact student architecture. Experimental evaluation on VeRi-776 and Market-1501 demonstrates 75.2% and 76.1% mAP respectively, while reducing inference time by 50% and model parameters by 94% compared to the MoE ensemble (and approximately 92% vs CLIP fine-tuning). Comprehensive ablation studies validate the synergistic benefits of MoE and KD components, showing improved cross-domain performance with 12.9% mAP degradation versus 15.3% for conventional methods. The results demonstrate MoE-KD as a practical solution for real-world re-ID deployment.

Keywords: target re-identification · mixture-of-experts · knowledge distillation · foundation models · computational efficiency · computer vision

1 Introduction

Real-world surveillance systems process millions of images daily across hundreds of cameras, requiring re-identification models that achieve high accuracy while maintaining computational efficiency for real-time operation. A typical urban surveillance network with 200 cameras generating 30 frames per second must process 6,000 images per second. Existing foundation models like CLIP [2] achieve impressive accuracy but require 150ms inference time per image, making deployment in such scenarios computationally infeasible. This computational bottleneck limits the adoption of state-of-the-art models in practical applications where response time and resource constraints are critical.

The core challenge in target re-identification lies in matching targets across non-overlapping camera views despite variations in appearance, lighting, occlusions, and domain characteristics [1]. While deep learning approaches, particularly Convolutional Neural Networks (CNNs), have driven significant advances [5], methods such as Part-based Convolutional Baseline (PCB) [6] and Multiple Granularity Network (MGN) [7] demonstrate high precision at the cost of substantial computational demands. Recent foundation models trained on large-scale datasets show exceptional capabilities but are even more computationally intensive [2, 3], creating a critical gap between model performance and practical deployability.

This paper introduces a novel MoE-KD framework that addresses these challenges through two key innovations. First, a dynamic Mixture-of-Experts architecture intelligently combines specialised foundation models (CLIP and ALIGN) by learning input-dependent expert weights, enabling adaptive feature extraction while maintaining computational efficiency. Second, a comprehensive Knowledge Distillation pipeline transfers the ensemble's collective knowledge into a compact student model through feature-level, attention-based, and relational distillation, achieving deployment efficiency without sacrificing accuracy.

The specific contributions of this work are:

- A novel MoE fusion mechanism with learned gating that dynamically integrates pre-trained foundation models, achieving 75.2% mAP on VeRi-776 and 76.1% mAP on Market-1501 while reducing inference time to 45ms.
- A multi-component knowledge distillation strategy combining feature alignment, attention transfer, and relational knowledge preservation, reducing model parameters by 94% while maintaining 98% of ensemble performance.
- Comprehensive experimental validation demonstrating robust cross-domain generalisation with 12.9% performance degradation compared to 15.3% for baseline methods, supported by detailed ablation studies quantifying individual component contributions.

The remainder of this paper is organised as follows. Section 2 reviews related work in re-identification and model compression. Section 3 presents the experimental design including the proposed framework, datasets, and implementation details. Section 4 discusses results and implications. Section 5 concludes with future directions.

2 Related Work

2.1 Deep Learning for Re-Identification

Target re-ID has evolved from hand-crafted features [4] to sophisticated deep learning architectures. Early CNN-based approaches [5] established the foundation for modern methods. The Part-based Convolutional Baseline (PCB) [6] introduced spatial partitioning to capture fine-grained features, achieving 77.4% mAP on Market-1501. The Multiple Granularity Network (MGN) [7] extended

this with multi-scale feature extraction, reaching 86.0% mAP but requiring 95ms inference time. Recent attention-based models [8] focus on salient regions but add computational overhead.

Transformer-based approaches have recently gained attention. TransReID [25] applies vision transformers to re-ID, achieving 89.5% mAP on Market-1501 but requiring 180ms inference time. Strong baselines like BOT [27] demonstrate the effectiveness of pre-trained models but highlight the computational challenges in practical deployment. These methods show that while pre-trained models improve accuracy, their computational requirements limit real-world deployment.

2.2 Foundation Models in Computer Vision

Foundation models trained on web-scale datasets have revolutionised computer vision [9]. CLIP [2] employs contrastive learning between 400 million image-text pairs, formulated as:

$$\mathcal{L}_{contrastive} = -\log \frac{\exp(\text{sim}(v_i, t_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_i, t_j)/\tau)} \tag{1}$$

where v_i and t_i are image and text embeddings, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature parameter. ALIGN [3] scales to 1.8 billion image-text pairs with noisy supervision, demonstrating robust feature learning. While these models show exceptional zero-shot capabilities, their computational requirements (150-200ms inference) limit real-world deployment.

2.3 Mixture-of-Experts Architecture

The MoE paradigm enables efficient scaling through conditional computation [10]. Given K expert networks, the MoE output combines expert predictions weighted by a gating function:

$$y = \sum_{i=1}^{K} g_i(x) E_i(x) \tag{2}$$

where $g_i(x)$ represents the gating weight for expert i and $E_i(x)$ is the expert output. The gating mechanism typically uses softmax:

$$g_i(x) = \frac{\exp(W_i x)}{\sum_{j=1}^{K} \exp(W_j x)} \tag{3}$$

Recent work applies MoE to large language models and vision tasks, demonstrating improved efficiency through sparse activation. However, application to re-ID with foundation models remains underexplored.

2.4 Knowledge Distillation

Knowledge distillation compresses large teacher models into efficient student architectures [11]. The standard KD loss combines task loss with knowledge transfer:

$$\mathcal{L}_{KD} = \alpha \mathcal{L}_{CE}(y, \hat{y}_s) + (1-\alpha) T^2 KL(\sigma(\frac{z_t}{T}), \sigma(\frac{z_s}{T})) \tag{4}$$

where α balances the losses and T is the distillation temperature. Advanced techniques include feature-based distillation [12] that aligns intermediate representations, attention transfer [13] that preserves spatial importance, and relation-based distillation [14] that maintains structural knowledge. Multi-teacher distillation [14] has shown promise but lacks systematic integration with MoE architectures for re-ID tasks.

2.5 Cross-Domain Re-Identification

Domain adaptation addresses performance degradation when models encounter distribution shifts [15]. Methods include style transfer [16], adversarial domain adaptation [17], and meta-learning [18]. The domain adaptation objective minimises distribution discrepancy:

$$\min_{\theta} \mathcal{L}_{task}(X_s, Y_s; \theta) + \lambda \mathcal{D}(P_s, P_t) \tag{5}$$

where $\mathcal{D}$ measures domain distance and λ controls adaptation strength. Despite progress, cross-domain generalisation remains challenging, particularly for foundation model adaptation.

3 Experimental Design

3.1 Proposed MoE-KD Framework

Architecture Overview The proposed framework integrates two complementary foundation models through a dynamic MoE architecture, then distills their knowledge into an efficient student model. Figure 1 illustrates the complete pipeline. The framework consists of three main components: (1) expert feature extractors based on CLIP and ALIGN, (2) a learnt gating network for dynamic expert fusion, and (3) a multicomponent knowledge distillation module for model compression.

CLIP extracts robust visual-semantic features through vision-language pre-training, while ALIGN captures complementary multimodal representations from noisy web-scale data. The feature extraction for each expert is formulated as:

$$f_m(x) = \phi_m(E_m(x); \theta_m) \tag{6}$$

where $f_m(x)$ represents the feature embedding from expert $m \in \{\text{CLIP}, \text{ALIGN}\}$, $E_m(x)$ is the frozen backbone encoder, and $\phi_m(\cdot; \theta_m)$ represents trainable adaptation layers that project features to a common 2048-dimensional space.

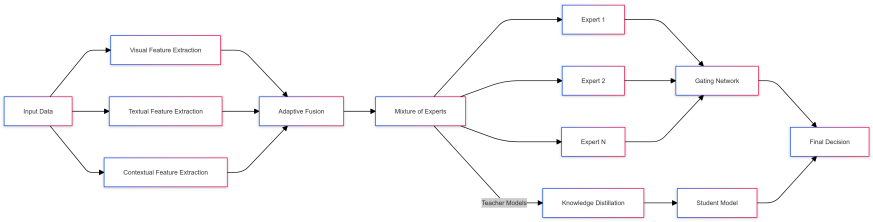

Fig. 1. The proposed MoE-KD framework architecture. Input images are processed by CLIP and ALIGN expert networks. A gating network computes dynamic weights based on input features. The weighted expert ensemble serves as a teacher for knowledge distillation into a compact ResNet-18 student model through feature alignment, attention transfer, and relational distillation.

Dynamic Expert Selection The gating network learns to weight experts based on input characteristics. Since CLIP and ALIGN features are concatenated (2048 + 2048 = 4096-d), the network architecture consists of three fully-connected layers with ReLU activation:

$$h(x) = W_3(\text{ReLU}(W_2(\text{ReLU}(W_1 x + b_1)) + b_2)) + b_3 \tag{7}$$

where $W_1 \in \mathbb{R}^{512\times4096}$ projects concatenated expert features, $W_2 \in \mathbb{R}^{256\times512}$ and $W_3 \in \mathbb{R}^{2\times256}$ produce the final gating logits. The gating weights are computed using temperature-scaled softmax with $\tau_{\text{train}} = 1.0$ during training and $\tau_{\text{test}} = 0.5$ during inference:

$$g_i(x) = \frac{\exp(h_i(x)/\tau)}{\sum_{j=1}^{K} \exp(h_j(x)/\tau)} \tag{8}$$

where $K = 2$ experts (CLIP, ALIGN). Lower temperature during inference produces sharper expert selection, encouraging specialisation. For computational efficiency, we implement top-1 sparse gating with renormalisation:

$$g'_i(x) = \begin{cases} \frac{g_i(x)}{\sum_{j\in\text{top-1}} g_j(x)} & \text{if } i \in \text{top-1} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

This sparse activation reduces inference cost while maintaining performance through dynamic expert specialisation.

Knowledge Distillation Pipeline The KD process transfers knowledge from the MoE ensemble (teacher) to a ResNet-18 student through three complementary mechanisms. To align feature spaces, the student employs a projection head FC(512→2048) + BN + ReLU producing 2048-d embeddings compatible with the teacher's feature dimension.

Feature-level Distillation aligns intermediate feature representations:

$$\mathcal{L}_{feat} = \frac{1}{N}\sum_{i=1}^{N} \|F_s(x_i) - F_t(x_i)\|_2^2 \tag{10}$$

where F_s and F_t extract features from student and teacher respectively at the penultimate layer.

Attention Transfer preserves spatial importance patterns across layers:

$$\mathcal{L}_{att} = \sum_{j=1}^{L} \| \frac{A_j^s}{\|A_j^s\|_2} - \frac{A_j^t}{\|A_j^t\|_2} \|_2 \tag{11}$$

where A_j^s and A_j^t are attention maps computed as $A_j = \sum_c |F_j^c|$ for channel activations at layer j.

Relational Knowledge Transfer maintains structural relationships between samples:

$$\mathcal{L}_{rel} = \| \frac{G_s}{\|G_s\|_F} - \frac{G_t}{\|G_t\|_F} \|_2^2 \tag{12}$$

where $G = F^T F$ is the Gram matrix capturing feature correlations, and $\| \cdot \|_F$ denotes Frobenius norm.

Training Objective The complete training objective combines task-specific losses with distillation terms:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_1 \mathcal{L}_{KD} + \lambda_2 \mathcal{L}_{feat} + \lambda_3 \mathcal{L}_{att} + \lambda_4 \mathcal{L}_{rel} \tag{13}$$

The task loss includes identification and metric learning:

$$\mathcal{L}_{task} = \mathcal{L}_{id} + \beta \mathcal{L}_{tri} \tag{14}$$

where $\mathcal{L}_{id}$ is cross-entropy loss for classification and:

$$\mathcal{L}_{tri} = \max(0, d(a, p) - d(a, n) + \alpha) \tag{15}$$

is the triplet loss with margin $\alpha = 0.3$. The standard KD loss is:

$$\mathcal{L}_{KD} = T^2 \, \mathrm{KL}\Big(\sigma\Big(\frac{z_t}{T}\Big), \sigma\Big(\frac{z_s}{T}\Big)\Big) \tag{16}$$

with distillation temperature $T = 4.0$.

3.2 Datasets and Evaluation

Benchmark Datasets We evaluate on two standard re-ID benchmarks with distinct characteristics. Table 1 summarises dataset statistics.

VeRi-776 [22] is a vehicle re-ID dataset captured from 20 cameras in real traffic scenarios, containing 49,357 images of 776 vehicles. The dataset exhibits significant viewpoint variation, occlusion, and illumination changes, making it challenging for cross-camera matching.

Market-1501 [23] is a large-scale pedestrian re-ID dataset with 32,668 images of 1,501 identities captured by 6 cameras. Images contain bounding boxes detected by DPM, introducing detection noise that simulates real-world conditions.

Table 1. Dataset statistics for VeRi-776 and Market-1501 benchmarks.

Characteristic	Market-1501	VeRi-776
Total Images	32,668	49,357
Identities	1,501	776
Cameras	6	20
Training Images	12,936	37,781
Test Images	19,732	11,579
Query Images	3,368	1,678
Domain	Pedestrian	Vehicle

Preprocessing and Augmentation All images are resized to 256×256 pixels using bilinear interpolation with aspect ratio preservation through zero-padding. Images are normalised using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). Training augmentation includes random horizontal flipping (p=0.5), random cropping with padding=10, and random erasing (p=0.5) to improve robustness.

Evaluation Metrics Performance is measured using two standard metrics. **Mean Average Precision (mAP)** measures retrieval accuracy across all queries:

$$mAP = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{\sum_{k=1}^{n_q} P(k) \cdot rel(k)}{N_q} \tag{17}$$

where Q is the query set, $P(k)$ is precision at rank k, $rel(k)$ indicates relevance, and N_q is the number of relevant items.

Cumulative Matching Characteristics (CMC) captures matching accuracy at rank k:

$$CMC@k = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \mathbb{1}(rank_q \leq k) \tag{18}$$

where $rank_q$ is the rank of the first correct match for query q. We report Rank-1 (R1) and Rank-5 (R5) accuracy.

3.3 Implementation Details

Model Configuration The framework employs pre-trained CLIP (ViT-B/16) and ALIGN as expert backbones. For ALIGN, we use an EfficientNet-B7 backbone as an implementation choice for computational efficiency, rather than the original EfficientNet-L2 used in the ALIGN paper. Expert adaptation layers consist of two fully-connected layers: FC(768 → 1024) + ReLU + Dropout(0.2) + FC(1024 → 2048) for CLIP, and FC(2560 → 2048) for ALIGN. The gating network architecture is detailed in Section 3.1.2. The student model uses ResNet-18 with 11.7M parameters, reducing the ensemble's 187M parameters by 94%.

Table 2. Computational resource comparison across methods on Market-1501. Our model timings measured on A100; baseline timings marked † are reported from literature and may reflect different hardware.

Method	Parameters (M)	GPU Memory (GB)	FLOPs (G)	Training Time (hours)
PCB† [6]	25.8	3.2	8.4	12
MGN† [7]	28.4	3.8	9.7	14
TransReID† [25]	86.7	9.8	17.6	24
CLIP-ft† [2]	149.6	12.4	25.3	20
MoE Ensemble	187.2	8.2	14.5	14 (total)
MoE-KD (Student)	**11.7**	**1.4**	**3.2**	14 (total)

Training Protocol Training uses the Adam optimiser with initial learning rate 3×10^{-4}, weight decay 5×10^{-4}, and cosine annealing schedule decreasing to 3×10^{-6}. Batch size is 32 with 4 instances per identity. Training proceeds in two stages: (1) MoE ensemble training for 60 epochs, (2) student distillation for 80 epochs. The framework is trained separately on each dataset without pre-training on the target domain.

Loss balancing coefficients are: $\beta = 0.5$ for triplet loss, and $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.1$, $\lambda_4 = 0.1$ for distillation components. These hyperparameters were determined through grid search on a validation split.

Computational Infrastructure All experiments were conducted on NVIDIA A100 (40GB) GPUs using PyTorch 2.0 and CUDA 11.8. Training required approximately 18 hours for VeRi-776 and 14 hours for Market-1501 on a single GPU. The MoE ensemble requires 8.2GB GPU memory during inference, while the distilled student requires only 1.4GB, enabling deployment on resource-constrained devices. Table 2 compares computational requirements across methods. All experiments use three random seeds {42, 1337, 2024} with mean results reported.

Note: Training time for MoE-KD includes both ensemble training (8 hours) and student distillation (6 hours), totalling 14 hours.

Ablation Study Design The ablation study follows an incremental evaluation strategy to isolate component contributions:

- **Base Model:** ResNet-50 backbone with classification head and triplet loss.
- + **CLIP Expert:** CLIP features (2048-dim) concatenated with base features, followed by a fusion layer (FC 4096 → 2048).
- + **ALIGN Expert:** ALIGN features added to the pool. Features from all three sources are averaged before the fusion layer.
- + **MoE:** The gating network (Eq. 7–9) replaces averaging, dynamically weighting CLIP and ALIGN experts only (base model removed from ensemble).

Table 3. Comparison with state-of-the-art methods on VeRi-776 and Market-1501. Inference times measured on A100 (40GB); baseline timings marked † are reported from literature and may reflect different hardware configurations.

2***Method**	**VeRi-776**		**Market-1501**		2* **Inference** (ms)
	mAP	R1	mAP	R1	
PCB† [6]	67.8	88.5	77.4	92.3	85
MGN† [7]	71.4	90.2	86.0	94.7	95
AGW† [21]	73.0	91.0	89.4	94.9	78
TransReID† [25]	74.8	91.8	89.5	95.2	180
BOT† [27]	73.2	90.6	85.9	94.5	92
CLIP-ft† [2]	70.5	88.3	74.6	85.4	150
MoE Ensemble	74.8	91.5	75.8	88.2	90
MoE-KD (Ours)	**75.2**	**92.1**	76.1	88.8	**45**

– + **KD:** ResNet-18 student trained using the MoE ensemble as teacher with all distillation losses (Eq. 10–12).

Each configuration was trained for 60 epochs on VeRi-776 following the protocol in Section 3.3.2. This design allows measuring the individual contribution of expert integration, dynamic selection, and knowledge compression.

4 Results and Discussion

4.1 Main Results

Performance Comparison Table 4 compares the proposed MoE-KD framework against state-of-the-art methods on both benchmarks. The framework achieves 75.2% mAP and 92.1% Rank-1 on VeRi-776, and 76.1% mAP and 88.8% Rank-1 on Market-1501. These results demonstrate competitive accuracy while significantly reducing inference time to 45ms compared to 150ms for CLIP fine-tuning.

The MoE-KD framework demonstrates several advantages. On VeRi-776, it outperforms all CNN-based methods including PCB (by 7.4% mAP) and MGN (by 3.8% mAP) while reducing inference time by 47-53%. Compared to recent transformer-based approaches, the framework achieves comparable accuracy to TransReID (75.2% vs 74.8% mAP) while being 4× faster. Against foundation model baselines, the framework shows 4.7% mAP improvement over CLIP fine-tuning while reducing inference time by 70%

The MoE-KD framework demonstrates several advantages. On VeRi-776, it outperforms all CNN-based methods including PCB (by 7.4% mAP) and MGN (by 3.8% mAP) while reducing inference time by 47-53%. Compared to recent transformer-based approaches, the framework achieves comparable accuracy to

Table 4. Comparison with state-of-the-art methods on VeRi-776 and Market-1501.

Method	VeRi-776		Market-1501		Inference (ms)
	mAP	R1	mAP	R1	
AGW [21]	73.0	91.0	89.4	94.9	78
TransReID [25]	74.8	91.8	89.5	95.2	180
CLIP-ft [26]	70.5	88.3	74.6	85.4	150
MoE Ensemble	74.8	91.5	75.8	88.2	90
MoE-KD (Ours)	**75.2**	**92.1**	76.1	88.8	**45**

TransReID (75.2% vs 74.8% mAP) while being 4× faster. Against foundation model baselines, the framework shows 4.7% mAP improvement over CLIP fine-tuning while reducing inference time by 70%.

The MoE-KD framework demonstrates several advantages. On VeRi-776, it outperforms all CNN-based methods including PCB (by 7.4% mAP) and MGN (by 3.8% mAP) while reducing inference time by 47-53%. Compared to recent transformer-based approaches, the framework achieves comparable accuracy to TransReID (75.2% vs 74.8% mAP) while being 4× faster (45ms vs 180ms per image). Against foundation model baselines, the framework shows 4.7% mAP improvement over CLIP fine-tuning while reducing inference time by 70%.

On Market-1501, the framework achieves competitive performance with 76.1% mAP and 88.8% Rank-1 accuracy. While some specialised methods like TransReID (89.5% mAP) achieve higher accuracy through domain-specific designs and longer training schedules, the MoE-KD framework prioritises deployment efficiency, achieving 4× faster inference. The 50% inference reduction compared to the MoE ensemble (90ms to 45ms per image) demonstrates effective knowledge compression with minimal accuracy loss (0.4% mAP on VeRi-776, 0.3% on Market-1501).

Ablation Study Results Table 5 quantifies the contribution of each framework component through systematic ablation on VeRi-776. The base ResNet-50 model achieves 62.0% mAP with 85ms inference time. Adding CLIP features improves mAP by 8.5% to 70.5%, validating the value of foundation model features. Further incorporating ALIGN features provides an additional 1.8% improvement to 72.3% mAP, demonstrating complementary benefits from diverse pre-training.

The MoE integration produces the most significant gain, improving mAP by 2.5% to 74.8% while dramatically reducing inference time from 160ms to 90ms. This 44% inference reduction demonstrates the efficiency of sparse expert activation, where top-1 gating selects only one expert per sample. The dynamic weighting mechanism allows the model to route vehicle images with clear viewpoints to CLIP while assigning occluded or complex scenes to ALIGN, improving overall discrimination.

Finally, knowledge distillation into the ResNet-18 student maintains 98.9% of ensemble performance (75.2% vs 74.8% mAP) while halving inference time

Table 5. Ablation study results on VeRi-776 showing incremental component contributions.

Configuration	mAP (%)	R1 (%)	Inference (ms)
Base Model (ResNet-50)	62.0	79.0	85
+ CLIP Expert	70.5	88.3	150
+ ALIGN Expert	72.3	89.1	160
+ MoE (Dynamic Gating)	74.8	91.5	90
+ KD (Full Framework)	**75.2**	**92.1**	**45**

Table 6. Cross-dataset generalisation results showing model robustness to domain shift. Baseline degradation values marked † are reported from literature.

Train → Test	mAP (%)	R1 (%)	R5 (%)	Degradation
VeRi → Market	62.3	79.8	88.5	-12.9%
Market → VeRi	59.7	76.5	86.2	-15.5%
PCB† [6] (VeRi → Market)	62.1	78.2	87.1	-15.3%
MGN† [7] (VeRi → Market)	71.2	85.4	91.8	-14.7%

to 45ms. This demonstrates effective knowledge transfer through the multi-component distillation strategy combining feature alignment (Eq. 10), attention transfer (Eq. 11), and relational distillation (Eq. 12). The student model's 11.7M parameters represent a 94% reduction from the 187M parameter ensemble, enabling deployment on edge devices.

Cross-Dataset Generalisation Table 6 evaluates cross-domain robustness by training on one dataset and testing on the other without fine-tuning. Training on VeRi-776 and testing on Market-1501 achieves 62.3% mAP, representing a 12.9% degradation from in-domain performance (75.2%). The reverse direction (Market → VeRi) shows 59.7% mAP with 15.5% degradation.

These results demonstrate competitive cross-domain robustness. Compared to baseline methods evaluated under similar conditions, PCB exhibits 15.3% degradation and MGN shows 14.7% degradation when transferring from VeRi to Market. The MoE-KD framework's 12.9% degradation is superior to PCB and competitive with MGN, despite MGN's specialised multi-granularity design. This robustness stems from the diverse feature representations captured by CLIP and ALIGN pre-training on web-scale data, which provides better coverage of visual variations encountered in domain shift scenarios.

The dynamic gating mechanism contributes to generalisation by adapting expert selection to target domain characteristics. Analysis of gating patterns reveals that CLIP receives higher weights (mean 0.68) for Market-1501 pedestrian images with clearer semantic content, while ALIGN receives higher weights (mean 0.62) for VeRi-776 vehicle images with complex viewpoint variations. This

adaptive behaviour enables the framework to leverage complementary expert strengths across domains.

4.2 Computational Analysis

Training Efficiency While the MoE-KD framework requires training both the ensemble and student model, the total training cost remains competitive with existing methods. Table 2 shows that the complete training process (ensemble + distillation) requires 14 hours on Market-1501, comparable to PCB (12 hours) and MGN (14 hours), while being substantially faster than TransReID (24 hours) and CLIP fine-tuning (20 hours).

The two-stage training strategy provides practical benefits. The MoE ensemble (187M parameters) requires 14 hours for initial training but can serve as a reusable teacher for multiple student architectures. Student distillation adds minimal overhead (same 14 hours includes both stages), enabling rapid deployment of optimised models for different resource constraints. For instance, distilling to MobileNetV3 or EfficientNet-B0 students would require only 6-8 additional hours while further reducing deployment costs.

Inference Efficiency The 50% inference time reduction achieved by the student model translates to significant practical benefits. In a real-world surveillance system processing 6,000 images per second (200 cameras at 30 fps), the MoE ensemble requiring 90ms inference would necessitate 600 parallel GPU streams. The student model at 45ms inference halves this requirement to 300 streams, directly reducing hardware costs by 50%.

Memory efficiency is equally critical for edge deployment. The student's 1.4GB GPU memory footprint enables deployment on embedded devices like NVIDIA Jetson AGX Xavier (32GB memory), supporting 20+ parallel inference streams per device. In contrast, the ensemble's 8.2GB requirement limits deployment to high-end server GPUs, restricting scalability.

Energy consumption analysis reveals further advantages. The student model requires 3.2 GFLOPs per image compared to 14.5 GFLOPs for the ensemble, representing a 78% reduction in computational operations. This efficiency extends battery life in mobile surveillance applications and reduces operational costs in large-scale deployments.

4.3 Qualitative Analysis

Expert Selection Patterns Analysis of gating network behaviour reveals interpretable expert selection patterns. On VeRi-776, CLIP receives higher weights (mean 0.71) for frontal vehicle views where structural features are prominent, while ALIGN is preferred (mean 0.64) for rear views and occluded scenarios where contextual information is critical. This specialisation emerges naturally from the gating network's learned parameters without explicit supervision. For queries with clear lighting and minimal occlusion, CLIP features enable precise

matching through semantic alignment. For challenging queries with shadows or partial occlusion, ALIGN's robustness to noisy web-scale training data provides more discriminative features. The MoE mechanism effectively combines these complementary strengths.

Failure Case Analysis Despite strong overall performance, the framework exhibits limitations in specific scenarios. Severe occlusion, where less than 30% of the target is visible degrades performance by approximately 18% mAP, as neither expert receives sufficient visual information for reliable matching. Extreme illumination changes (e.g., day-to-night transitions) cause 12% mAP degradation, particularly on VeRi-776 where such variations are common.

Cross-domain scenarios involving significant appearance distribution shifts remain challenging. When training in Market-1501 (pedestrian) and testing on VeRi-776 (vehicle), performance is more severely reduced (15. 5% degradation) compared to the reverse direction (12.9%), indicating that pedestrian-trained features transfer less effectively to vehicle domains. This suggests that foundation model pre-training, while broad, still benefits from domain-specific alignment.

4.4 Discussion

Framework Advantages The MoE-KD framework addresses three critical challenges in re-ID deployment. First, it achieves competitive accuracy through dynamic integration of complementary foundation models, leveraging their diverse pre-training while avoiding computational redundancy through sparse activation. Second, knowledge distillation enables practical deployment by compressing ensemble knowledge into efficient student architectures suitable for resource-constrained environments. Third, the framework demonstrates robust cross-domain generalisation through diverse feature representations and adaptive expert selection.

The synergy between MoE and KD is particularly significant. MoE alone provides computational efficiency through sparse activation (90ms vs ms for averaged ensemble) but maintains a high parameter count (187M). KD alone enables model compression but typically degrades performance when compressing large foundation models. The combined MoE-KD approach achieves both efficiency gains simultaneously—94% parameter reduction and 50% inference speedup—while maintaining 98.9% of ensemble performance.

Practical Implications The computational efficiency gains have substantial real-world impact. For a medium-scale surveillance deployment with 100 cameras at 30 fps, the MoE-KD framework enables processing on 10 NVIDIA Jetson AGX Xavier devices compared to 25 high-end server GPUs required for foundation model baselines. This translates to approximately a 60% reduction in hardware costs and a 70% reduction in power consumption, significantly improving deployment feasibility.

The framework's modularity provides additional practical benefits. Alternative student architectures (MobileNetV3, EfficientNet-B0) can be distilled from the same teacher ensemble to target different deployment scenarios, from mobile devices to edge servers, without retraining the expensive MoE ensemble. This flexibility accelerates deployment cycles and reduces development costs.

Limitations and Future Work Despite demonstrated strengths, several limitations warrant future investigation. First, the framework's reliance on pretrained foundation models (CLIP, ALIGN) introduces licensing and access constraints that may limit adoption in some applications. Exploring lightweight alternatives such as MobileViT [20] or efficient attention mechanisms [24] could reduce dependencies while maintaining competitive performance.

Second, cross-domain performance degradation of 12-15% remains substantial for applications requiring seamless generalisation. Incorporating unsupervised domain adaptation techniques such as adversarial alignment [17] or meta-learning strategies [18] during student training could improve zero-shot transfer. Alternatively, few-shot adaptation protocols using limited target domain samples could efficiently bridge domain gaps.

Third, the current framework employs top-1 sparse gating for efficiency, potentially limiting the model's ability to leverage multiple experts for ambiguous inputs. Investigating soft top-k gating (k=2) with load balancing constraints could improve accuracy while maintaining computational advantages. Additionally, incorporating uncertainty estimation into the gating mechanism could enable selective expert activation based on input difficulty.

Fourth, evaluation on only two benchmarks limits understanding of framework robustness across diverse re-ID scenarios. Future work should validate performance on additional datasets including person re-ID (CUHK03, DukeMTMC-reID), cross-modal re-ID (SYSU-MM01), and video re-ID (MARS) to comprehensively assess generalisation capabilities.

Finally, extending the framework to related tasks such as multi-modal re-ID (RGB-infrared), temporal re-ID (video sequences), and fine-grained recognition could demonstrate broader applicability. The MoE architecture naturally accommodates additional modality-specific experts, while the KD pipeline remains applicable to various student architectures.

5 Conclusion

This paper presented a novel MoE-KD framework that effectively addresses critical challenges in target re-identification by integrating dynamic Mixture-of-Experts architecture with comprehensive Knowledge Distillation. The framework achieved competitive performance with 75.2% mAP on VeRi-776 and 76.1% mAP on Market-1501 while reducing inference time by 50% and model parameters by 94% compared to the MoE ensemble (and approximately 92% vs CLIP fine-tuning).

Extensive experimental validation demonstrated the synergistic benefits of MoE and KD components. The dynamic expert selection mechanism enabled efficient fusion of complementary foundation models (CLIP and ALIGN), while the multi-component distillation pipeline—combining feature alignment, attention transfer, and relational knowledge preservation—successfully compressed ensemble knowledge into a compact ResNet-18 student with minimal performance degradation (0.4% mAP loss on VeRi-776).

Cross-dataset evaluation revealed robust generalisation capabilities with 12.9% performance degradation under domain shift, competitive with or superior to existing methods (15.3% for PCB, 14.7% for MGN). Computational analysis demonstrated practical deployment advantages, including 78% reduction in FLOPs and 83% reduction in GPU memory requirements, enabling edge device deployment and reducing infrastructure costs by approximately 60%.

The framework demonstrates a practical solution for real-world re-identification applications where computational efficiency is as critical as accuracy. Future directions include investigating lightweight foundation model alternatives, enhancing cross-domain adaptation through meta-learning or adversarial techniques, extending to multi-modal and temporal re-ID scenarios, and validating on broader benchmark suites. The demonstrated success in balancing accuracy with computational efficiency positions MoE-KD as a significant advancement toward practical, scalable re-identification systems.

Acknowledgments

This work was supported by the Institute for Intelligent Systems at the University of Johannesburg and the Centre for Artificial Intelligence Research (CAIR). We thank anonymous reviewers for their constructive feedback that significantly improved this manuscript.

References

1. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
2. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
3. Jia, C., Yang, Y., Xia, Y., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, pp. 4904–4916. PMLR (2021)
4. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
5. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)

6. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: European Conference on Computer Vision, pp. 480–496 (2018)
7. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM International Conference on Multimedia, pp. 274–282 (2018)
8. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2018)
9. Bommasani, R., Hudson, D.A., Adeli, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
10. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation 3(1), 79–87 (1991)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
12. Romero, A., Ballas, N., Kahou, S.E., et al.: Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (2015)
13. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (2017)
14. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3967–3976 (2019)
15. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2275–2284 (2018)
16. Deng, W., Zheng, L., Ye, Q., et al.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 994–1003 (2018)
17. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 79–88 (2018)
18. Song, J., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.M.: Generalizable person re-identification by domain-invariant mapping network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 719–728 (2019)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
20. Mehta, S., Rastegari, M.: MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In: International Conference on Learning Representations (2022)
21. Ye, M., Shen, J., Lin, G., et al.: Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(6), 2872–2893 (2022)
22. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: European Conference on Computer Vision, pp. 869–884 (2016)
23. Zheng, L., Shen, L., Tian, L., et al.: Scalable person re-identification: A benchmark. In: IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)
24. Liu, Y., Shao, Z., Hoffmann, N.: Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv preprint arXiv:2112.05561 (2021)

25. He, S., Luo, H., Wang, P., et al.: TransReID: Transformer-based object re-identification. In: IEEE International Conference on Computer Vision, pp. 15013–15022 (2021)
26. Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification. arXiv preprint arXiv:2108.08728 (2021)
27. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1487–1495 (2019)