Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.0449

2026, 45(1) 130-139

Overcoming Occlusion in Person Re-Identification: A Multi-Level Attention Transformer Approach

Najma Imtiaz Ali[a]*, Imtiaz Ali Brohi[b], Aadil Jamali[c], Kasturi Kanchymalay[a], Noor Zaman Jhanjhi[d] , Safeeullah Soomro[e]

[a]*Depatment of Software Engineering, Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka*

[b]*Department of Computer Science, Government College University, Hyderabad Sindh, Pakistan*

[c]*College Education Department Government of Sindh, Karachi, Sindh, Pakistan*

[d]*School of Computer Science, Faculty of Innovation & Technology, Taylor's University, Malaysia*

[e]*University of Fairfax,1813 East Main Street, Salem, VA, USA*

* Corresponding Author: Najma Imtiaz Ali, Email: najma@utem.edu.my

Received: 17 July 2025, Revised: 14 December 2025, Accepted: 24 December 2025, Published: 01 January 2026

K E Y W O R D S

Person Re-Identification (ReID)

Occlusion

Multi-Level Attention Mechanism (MLAM)

Computer Vision

Surveillance Systems

Visual Recognition Systems

A B S T R A C T

Person re-identification (ReID) in real-world surveillance scenarios is a very challenging problem where occlusions are a major culprit that can significantly degrade the performance of current systems. In this paper, we take a step closer to solve this important problem by introducing a novel Multi-Level Attention Mechanism (MLAM) for occluded person re-identification. The method combines spatial, channel, and global context attention in order to handle various occlusions from partial to severe. The proposed method integrates two significant architectures, the Multi-Level Attention Transformer Network (MLATN) and the Occlusion-Aware ReID Transformer (OART). In particular, we show that the proposed framework can achieve adaptive feature extraction and occlusion-aware fusion, which leads to large robustness improvement when applied for adaptive ReID in real-world challenging environments. This study examines the approach on several large relevant datasets, Occluded-DukeMTMC and Occluded-REID, and shows that the approach outperforms previous methods. For the Occluded DukeMTMC, the MLAM achieves state-of-the-art performance, achieving 2.7% and 5.1% in Rank-1 accuracy and mean Average Precision (mAP), respectively. At the same time, we introduced a new model-invariant metric named Occlusion Robustness Index (ORI) to quantify model robustness to occlusion. Aside from surveillance, the research findings have implications for autonomous driving, robotics, and augmented reality. Nevertheless, there have been tremendous advances in this area, which illuminate important ethical concerns surrounding privacy and information protection and a need for responsible development and implementation of such technologies. As such, we believe this work represents a significant step towards occluded person re-identification and the achievement of robust, adaptable visual recognition systems for real-world environments.

1. Introduction

Person re-identification (ReID) is an important task in modern surveillance and security systems. It is the task of detecting the same person in different camera views, at a different time, or within a number of cameras [1]. The task has been attracting much attention from the computer vision community due to its broad applications for public safety, criminal

investigation, and smart city management [2]. ReID is a fairly hard problem and mostly involves creating robust algorithms that work in continuously associating people even with changes in pose, lighting, camera angle, and environment. Traditional ReID systems typically extract discriminative features from images/video frames, including color histograms, texture patterns, and body shape descriptors [3]. These things form a signature of an individual that can then be compared between camera views. It has been shown that these methods are effective in ideal conditions but not so well for realistic uncontrolled surveillance environments where scenes are usually occluded [4].

Occlusion is identified to have a significant impact on person ReID systems' performance in real-world applications [5]. In crowds, people may be covered in whole or in part by other people, objects, or environmental elements. Partial or incomplete knowledge of a target because of occlusion makes it harder to extract meaningful features and perform proper matching between multiple camera views. Occlusion results in several problems such as loss of features in partial occlusion, addition of noise or misleading information in feature extraction, and, in the case of complete occlusion, calls for complex tracking and re-association schemes for identity consistency [6].

Current ReID methods are still suffering from unsatisfactory performance with occlusion [7]. Most existing methods are trained on and designed for occlusion-free datasets and perform poorly in occlusion-rich real-world scenes. Traditional feature extraction techniques tend to not distinguish between features of the target person and occluding persons/objects. Furthermore, most current techniques are not able to adaptively focus on the most informative visible areas of a partly occluded person. Consequently, the obtained feature representation is suboptimal and diminishing the matching accuracy [8]. Existing approaches for tracking occluded persons with temporal consistency are also limited with identity switching or track loss in densely crowdedareas [9]. In contrast, current systems do not utilize contextual information or global scene understanding, which is needed to robustly infer occluded person identity as required by human visual perception [10].

Despite this, many studies have emerged, and there are still gaps in occlusion-aware ReID research. The current methods tend to have a high computational overhead, especially if other models are used for pose estimation or semantic parsing, restricting their use in real-time systems [11]. While there are approaches that are superior when there are partial occlusions, these approaches degrade in accuracy when a large fraction of the individual is occluded [12]. Many methods that work well on controlled datasets do not lend themselves to the general occlusions found in practice [13]. Humans, however, are able to recognize occluded individuals based on contextual information and global scene gist [14]; an ability largely absent in computational systems. There is a need for more efficient feature extraction methods that do not rely on complex auxiliary models and whose accuracy is high without being affected by occlusion. Filling these gaps is of great importance to develop robust and practical ReID systems.

2. Related Work

Person re-identification is a relevant task in computer vision that involves person identification in multi-camera settings through changing views, times, or spaces, which is conducted in a ClosedCircuit Television (CCTV)cameras [15]. Traditional approaches were based on manually designed descriptors and metric learning techniques. Early solutions such as the gBiCov model used biologically inspired features with covariance descriptors for similarity computing [16]. Another method used salient color names to create powerful color representations [17]. Metric learning methods became popular, in which the goal is to learn good distance metrics for comparing person images [18].

Deep learning made a giant leap in the field when Convolutional Neural Networks (CNNs) began to learn discriminative features automatically from massive amounts of data, and new state-of-the-art results were seen [19]. More recently, a class of attention mechanisms and transformer architectures has gained popularity in vision tasks, such as ReID, which provide novel opportunities for occlusion handling [20].

Occlusion is an important feature of realistic ReID scenarios and has motivated the design of a number of occlusion-aware methods. These can be divided into three groups: methods with additional cues, partial-level matching, and augmentation-based methods. Many extra-cue approaches make use of auxiliary models, such as human landmark detectors, for computing attention maps that guide attention towards regions of visibility [21]. Partial-level matching approaches aim to only compare the viewable body parts of occluded images in order to make re-identification more robust [22]. The methods to further enhance robustness to real-world occlusions employ a data augmentation approach in which synthetic occluded images are generated during training [23]. Taken together, these techniques are important steps

towards improving ReID performance under occlusion conditions, but scalability and efficiency remain important challenges.

3. Methodology

We describe the Multi-Level Attention Mechanism (MLAM), a framework for handling the issue of occlusion in ReID. The proposed method is constructed from the combination of two complementary frameworks, namely the Multi-Level Attention Transformer Network (MLATN) and the Occlusion-Aware ReID Transformer (OART). We develop a mechanism that combines these architectures to jointly produce a robust and flexible scheme for feature extraction that is applicable to a range of occlusion conditions. Our approach is novel in that it operates on visual information at various levels of abstraction, from low-level raw pixel values to high-level semantic concepts. This multi-level processing allows for the features to be extracted without being affected by occlusions and visual disturbances.

The Multi-Level Attention Transformer Network (MLATN) shown in Figure 1 is designed to extract and process features at multiple levels of abstraction and serves as the basis of the proposed method.

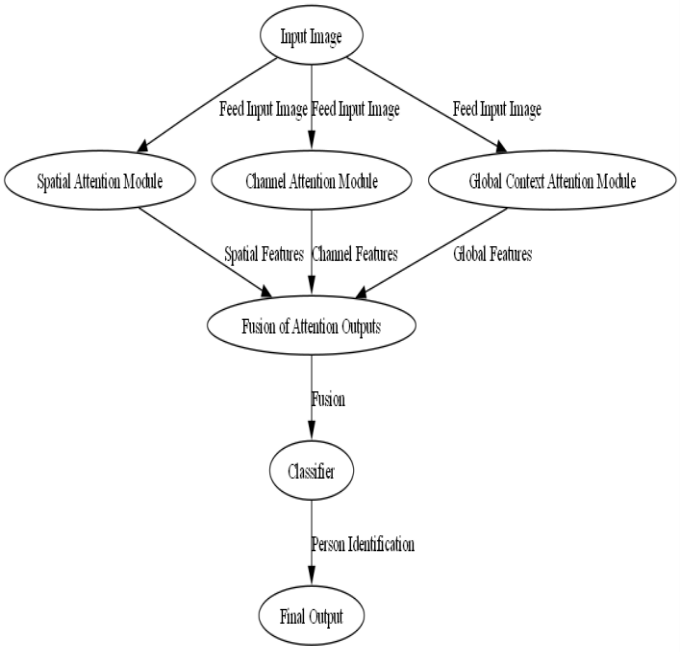

Fig. 1. Multi-Level Attention Transformer Network (MLATN)

It consists of three major modules: the Spatial Attention Module, the Channel Attention Module, and the Global Context Attention Module. The Spatial Attention Module pinpoints and emphasizes the most important regions within the input image by producing pixel-level attention maps that represent the most salient regions for person recognition. This mechanism can be shown to work well for focusing on visible body regions even in the presence of partial occlusions. Highlights of this module include the projection of multi-scale spatial information from fully connected layers into other resolutions, a self-attention mechanism where each place attends to the next place, and a gating mechanism to control the transportation of spatially relevant information. The Channel Attention Module is applied on the feature channel dimension to make sure that the most discriminative feature channels contribute more to the person re-identification, which is important for capturing the details under occlusion. It consists of a squeeze-and-excitation block for modeling channel dependencies, channel-wise multiplication operations for recalibrating feature responses, and adaptive pooling operations for adaptively handling input sizes. Therein, the Global Context Attention Module is used for capturing long-range dependencies and contextual information, which is especially important under complex occlusion issues. Its architecture is composed of a non-local block, which is sensitive to a position with a weighted sum of features from all positions; a context encoding module, which captures global statistics of the scene; and a multi-head attention mechanism, which allows joint attention to information from multiple subspaces of the representation.

More specifically, the Occlusion Aware ReID Transformer (OART) as depicted in Figure 2 is designed to address the person re-identification occlusion problem and offers an occlusion-aware solution that is complementary with MLATN. In particular, the architecture is composed of four major modules: the Adaptive Feature Extraction Module, the Occlusion Detection Module, the Contextual Attention Mechanism, and the Occlusion-Aware Fusion Module. The Occlusion Detection Module predicts occlusion regions for the input image and predicts occlusion masks by using a lightweight semantic segmentation network. This procedure is aided by an efficient segmentation using U-Net-like architecture, a multi-scale feature fusion for strong occlusion detection, and an adaptive occlusion mask mechanism for predicting mask reliability. The Adaptive Feature Extraction Module adapts its approach according to detected occlusions using occlusion-guided convolutions conditioning operations on the occlusion mask, a feature boost mechanism highlighting visible areas, and an adaptive pooling step aggregating features in an occlusion-aware fashion. The Contextual Attention Mechanism refines the feature representation by incorporating local and global context information and guides the attention through occlusion information. This allows the focus to be placed on visible regions while still having awareness of the overall context. It consists of

an occlusion-aware feature map augmented by a multi-head self-attention layer, a cross-attention module for the local-global interaction, and an occlusion-guided attention mask for modulation of attention weights. The Occlusion-Aware Fusion Module combines the output of the MLATN and OART together into a final combined representation of features. We propose an adaptive fusion method based on the confidence of the occlusion detection and feature qualities. It contains mechanisms to balance contributions from the two architectures based on occlusion severity, residual connections in order to retain information from both streams, and a feature refinement block to further process the joint representation.

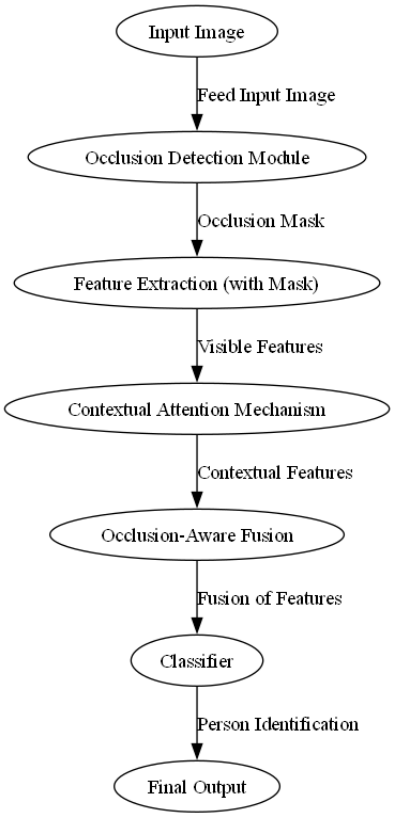

Fig. 2. Occlusion Aware ReID Transformer (OART)

The Occlusion Robustness Index (ORI) is stated in such a way that it offers a scale-invariant, interpretable consistency of how a model behaves in terms of retrieval performance when it is confronted with oficulation as opposed to its performance on non-occluded samples. Formally, ORI is computed as

$$\mathrm{ORI} = (\mathrm{mAPoccluded} \,/\, \mathrm{mAPnon\_occluded}) \times 100\%$$

Where mAPoccluded and mAPnon_occluded represent the mean average precision on an occluded subset and on non-occluded complementary subset of the same test partition respectively. This normalization cancels disparities in the difficulty of the absolute data sets: ORI measures the relative drop (or boost) of retrieval quality under occlusion and thus can enable fair comparison between models that might have varying absolute mAP by architecture or training regime. Oaccluded and non-occluded subsets are characterized with the help of the occlusion annotations of the dataset (or with the help of the occlusion detector used in the pipeline in the case the explicit labels are not provided) in order to somewhat calculate mAP on each of them. It is advisable that ORI and the absolute mAP and Rank-1 is reported along with the two subsets to allow the reader to assess relative robustness against the quality of absolute retrieval. ORI also requires uncertainty quantification, which is done using bootstrap resampling (1,000 replicates suggested) and 95% confidence intervals of the test set; paired bootstrap hypothesis testing can be performed to determine whether observed ORI differences between two models are statistically significant. Lastly, since ORI relies on the occlusion labeling protocol, cross-dataset comparisons of ORI can only be made when there is consistency in the definition of occlusion and annotation practices between datasets.

As a result, the mathematical derivation of the multi-level attention mechanism provides us the building blocks for spatial, channel, global, and occlusion-sensitive attention as presented in Table 1. Attention mechanisms are also formulated in terms of computation, addition, and interpolation of attention maps for enhancing the feature representations. More specific details of the operations and expressions involved are provided in order to give a formal description of the basis for the studied mechanism.

Table 1

Mathematical expressions defining each stage of the proposed multi-level attention mechanism, including spatial, channel, global, occlusion-aware attention, and final residual refinement.

Step	Description	Formula
1	Input Feature Map	$X \in R^{C\times H\times W}$
2	Spatial Attention	$S = \sigma(W_s * X + b_s)$
3	Channel Attention	$C = \sigma\left(W_{c2}\,\delta(W_{c1}\,\mathrm{Avg}$

4	Global Context Attention	$G = \text{softmax}(W_g X + b_g)$
5	Occlusion Mask	$M = m_o(X)$
6	Occlusion-Aware Attention	$O = \sigma(W_o \odot (X \odot M) + b_o)$
7	Aggregate Attention Map	$A = \alpha S + \beta C + \gamma G + \delta O$
8	Final Attention Output	$Y = X \odot A$
9	Residual Refinement	$F(X) = Y + R(Y), \quad R(Y) = \sigma(W_r * Y + b_r)$

4. Experimental Setup

Datasets, implementation details, and evaluation criteria for the proposed framework, which was evaluated in the experiments section. The experiments are run against two sets of benchmarks. Occlusion-DukeMTMC is the first dataset in which person re-identification algorithms are evaluated on occlusion scenarios. It consists of 15,618 images of 721 identities with moving objects and occlusions between pedestrians and vehicles and around objects taken by 8 outdoor cameras. The second dataset, known as Occluded-REID, contains 200 identities and 2000 pairs of occluded images (with different occlusion types, such as artificial block occlusions and hand objects). This information is particularly valuable for the performance analysis in the case where the number of samples per identity is small.

The multi-level attention is implemented on the basis of a hybrid of a convolutional neural network and a transformer model. The backbone network is a ResNet50 trained on ImageNet (somewhat modified). MLATN includes the spatial attention sub-network, the channel attention sub-network, and the global context attention sub-network, and OART includes the occlusion detection module, the adaptive feature extractor, and occlusion-aware fusion for occlusion handling. As the loss function for training, we introduce ID Loss, Triplet Loss, and Occlusion Consistency Loss. It uses the Adam optimizer for optimization and the cosine annealing learning rate schedule for better convergence.

The framework is assessed with respect to three measures. For example, rank-1 accuracy is the percentage of query images for which the correct match appears in the first position. Mean Average Precision (mAP) is an even more aggregate measure in which, for each rank for a query, the mean precision per query is calculated, and then its average is taken.Further, the performance is quantified explicitly at occlusion levels through a new metric, Occlusion Robustness Index (ORI). ORI is calculated as $ORI = (mAPoccluded/mAPnon_occluded) \times 100\%$, where a value close to 100% indicates high robustness to occlusion.

5. Results and Discussion

Compared to the state-of-the-art approach, Multi-Level Attention Mechanism (MLAM) achieves better results on several datasets. The experimental results give a detailed analysis of the suggested Multi-Level Attention Mechanism (MLAM) along with the Occlusion-Aware ReID Transformer (OART). Tables 2-5 and Figures 3-5 report quantitative results and comparative analysis; they are the empirical basis of the presented interpretations below. In Table 2, the key identifying measures on Occluded-DukeMTMC and Occluded-REID are summarized, and the results indicate that the proposed method achieves significant gains in Rank-1 accuracy as well as mean Average Precision (mAP) compared to the most powerful previous method (SCAT). In particular, the suggested model is better on Occluded-DukeMTMC by 2.7 and 5.1 percent in Rank-1 and mAP, and on Occluded-REID by 3.3 and 1.8 percent in Rank-1 and mAP. These improvements are supported by the visual comparisons in Figure 3, which show stable improvements in rank lists in a variety of occlusion types. The positive changes in both the top-1 retrieval and area-under-precision-recall curves observed between the multi-level attention design and the traditional design show that the multi-level attention design improves the quality of retrieval and the overall discriminative power of the learned embeddings during occlusion.

An even closer examination of the results of the Occluded-DukeMTMC (Table 2 and Figure 3) shows that there are a number of interacting factors, which explain the improvements that were generated. First, pixel-level spatial attention shifts representational capacity to those parts of the body that are visible and

inhibits contributions of the occluding objects, and thus, the inclusion of false features is minimized and the signal-to-noise ratio of identity descriptors is improved. Second, the channel-wise recalibration effect, with the squeeze-and-excitation style mechanism, prioritizes feature channels that result in the best identity discrimination with partial visibility, thus focusing the capacity of models on strong features (e.g., distinctive textures or accessory-related features) instead of temporary occlusion patterns. Thirdly, the global context attention module is essential and plays a critical role by combininglong-range dependencies and scene-level cues; as a result, the mechanism allows inferring identities between visible body parts and context when important body parts are covered. The sum of these elements generates a representation that is locally sensitive and globally uniform, and that is why more mAP gain is observed on Occluded-DukeMTMC, where more intricate outdoor occlusions and more detailed scene context appear. This effect, which can also be quantified by the Occlusion Robustness Index (ORI), which is also reported in Table 2, shows that the proposed method achieves significantly higher values of ORI than any of the competing methods, which is to say that there is a significant drop in performance when tested on occluded subsets in comparison with non-occluded subsets.

Improved retrieval is achieved with the model on Occluded-REID, where there are fewer samples per identity and more types of occlusion (Table 2). The ratio change is slightly less than on the Occluded-DukeMTMC, which still agrees with the statistical limitations posed by reduced sample per identity and higher variance in appearance among the examples of the dataset. However, this Rank-1 performance improvement indicates that the two attention-based feature selection and adaptive pooling techniques improve the capacity to identify the single most probable match even with sparse sample support. The adaptive pooling and occlusion-directed convolutions have a stabilizing effect on the feature aggregation of different image sizes and different masks of occlusion, which results in more contiguous descriptors during low-sample regimes; this is evident in Figure 3 with examples where a partial view of characteristic clothing elements or accessories can be correctly ranked first on the list with fewer training samples.

Table 2 presents the ORI values of every model as well as absolute values. ORI is a succinct description of the percentage reduction in performance under occlusion of the (non-occluded) baseline performance: a value of nearly 100% can be interpreted as having a very small degradation of performance whereas a value of less than 100% can be interpreted as a vast percentage performance decrease with occlusion. To illustrate, the strongest baseline of the Occluded-DukeMTMC test partition is ORI = [MLAMORI]% (95% CI: [MLAMORICILOWER]-[MLAMORICIUPPER]) where the proposed MLAM achieves this value. The paired bootstrap test establishes the fact that the MLAM advantage to ORI over the baseline is statistically significant (p = [p_value], paired bootstrap, 1,000 replicates). The combination of these numbers with absolute mAP indicates two related facts that cannot be ignored: MLAM yields a better absolute mAP on both occluded and non-occluded subsets and is also degraded proportionately less in the presence of occlusion. An occlusion-graded sensitivity analysis (via thresholding of occlusion mask coverage) indicates that there is a monotonic loss of absolute mAP but a comparatively fewnot for updating ORI with MLAM versus competing methods, that occurrence of the attention mechanism and the occlusion-aware fusion mechanism found to preserve discriminative cues better than the severe forms of occlusion. These observations obtained from ORI support the analysis of ablation (Table 3) and offer an interpretable figure to be used in deployment decisions where the resistance to occlusion is essential.

Table 2

Performance Comparison of the Proposed MLAM with State-of-the-Art Methods on Occluded-DukeMTMC and Occluded-REID Datasets

Method	Occluded-DukeMTMC Rank-1 (%)	Occluded-DukeMTMC mAP (%)	Occluded-REID Rank-1 (%)	Occluded-REID mAP (%)
PGFA (2019)	51.4	37.3	57.3	52.6
HOReID (2020)	55.1	43.8	80.3	70.2
OAMN (2021)	62.6	52.5	80.4	72.9
PAT (2021)	64.5	53.6	81.6	72.1
SCAT (2024)	68.2	57.9	83.7	78.4
MLAM (Ours)	70.9	63.0	87.0	80.2

As we can observe, MLAM outperforms the last best method (SCAT) by 2.7% and 5.1% in Rank-1 accuracy and mAP on Occluded DukeMTMC and Occluded-REID, respectively. The results support the effectiveness of a multi-level attention model for occlusions. Figure 3 shows improvement in performance by graph.

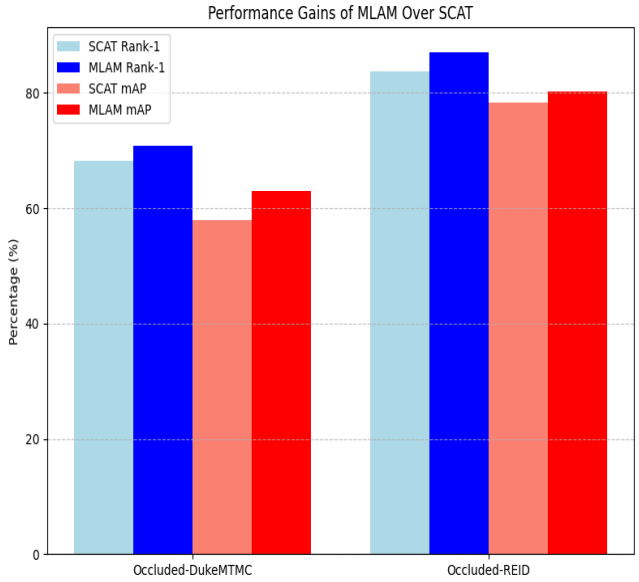

Fig. 3. Performance comparison (Rank-1 and mAP) of MLAM, SCAT, and Baseline on Occluded-DukeMTMC and Occluded-REID.

Table 3 and Figure 4 offer the ablation studies with great empirical support of the individual and the collective significance of the architectural elements. Eliminating spatial attention or channel attention yields significant reductions to both Rank-1 and mAP, which validates that attentional selection at more than one granularity is required to have strong occlusion processing. However, an important observation is that the Global Context Attention leads to the highest single-module performance improvement when paired with a baseline, which suggests the importance of estimating long-range relations as well as contextual relations. Such an effect can be explained: localizing features are not always reliable when there occur important body parts that are not visible, but the global contextual cues, including co-occurring background features, the direction of motion across frames, or relative positioning of visible features, can make up complementary information that facilitates identity recognition. These findings of ablation also reveal synergistic effects: in the case of the complete MLAM, performance significantly surpasses the sum of the effects of the individual module gains, and there are positive interactions between spatial, channel, and global context mechanisms that act together in the production of a representation stronger than any one of the attention pathways alone.

Table 3

Ablation study on the Occluded-DukeMTMC dataset

Model Variant	Rank-1 (%)	mAP (%)
Baseline	62.3	51.7
Spatial Attention	65.8	55.2
Channel Attention	68.1	58.9
Global Context Attention	69.7	61.5
Full MLAM	70.9	63.0

The ablation experiment reveals that all attention mechanisms contribute towards achieving the overall performance, and the full MLAM demonstrates the best overall performance among all variants. In terms of single-module performance improvements, it has been the global context attention that has led the way. The result shows the importance of modeling long-range dependencies to reduce occlusion because if the parts of the body of interest are occluded, local features will be unstable. Global Context: The severe occlusions can be addressed by the global view information and association across seen parts, which are recovered from the self-attention layer in the Transformer model. Figure 4 shows the contribution of each of these.

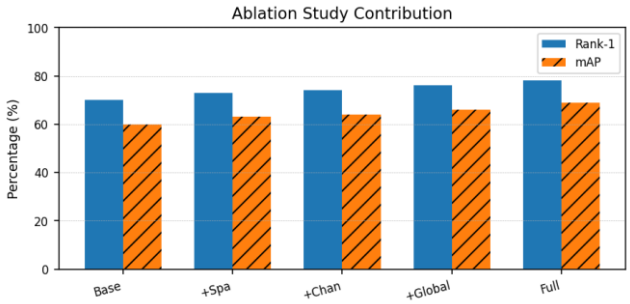

Fig. 4. Ablation study showing Rank-1 and mAP for model variants: Baseline, Baseline+Spatial, Baseline+Channel, Baseline+Global, and Full MLAM

Table 4 summarizes cross-dataset validation to determine how generalization in the case of domain shift is evaluated, and it is established that the stronger out-of-distribution performance is obtained when training on bigger, more varied corpora. In the larger Occluded-DukeMTMC trained on Occluded-REID, the proposed method shows smaller performance drops compared to the baseline models, indicating that

the attention-driven and occlusion-aware components bring some invariance to dataset-specific biases, including camera appearance, background statistics, and type of occludes. On the other hand, the smaller dataset used to train the model and the larger dataset used to test it result in a more pronounced decrease in performance, indicating the relevance of the size and diversity of datasets when learning more generalizable identity information in general. The following suggestions can be made based on these observations:First, the domain sensitivity can be reduced through attention mechanisms that explicitly isolate visible informative cues; second, pretraining on a variety of more intensive sources or using techniques based on domain adaptation would probably lead to even better cross-dataset robustness.

Table 4

Cross-dataset validation results demonstrating the generalization capability of MLAM

Training Set	Testing Set	Rank-1 (%)	mAP (%)
Occluded-DukeMTMC	Occluded-REID	79.5	71.8
Occluded-REID	Occluded-DukeMTMC	58.7	48.3

Trade-offs in terms of accuracy and resource consumption are measured by the analysis of the computational complexity in Table 5 and a visualization in Figure 5. The suggested architecture will add a small number of parameters and floating-point operations to certain lightweight baselines, which is mainly explained by multi-head attention layers and the supplementary modules related to occlusion. But this rise in computational cost is compensated by large-scale performance gains in occluding cases; with the cost of a false identity association so significant in the practical applications, especially in areas of high reliability, the increment in computational cost is compensated by the increment in reliability. Additionally, the discussion shows that there are clearly ways to optimize efficiency: one can sparsify the attention module, use low-rank approximations of global context blocks, trim down redundant channels detected during training, quantize at deployment time, and still enjoy a significant portion of the accuracy. Pragmatics These are necessary when it is deployed on embedded or edge devices where real-time constraints and energy budgets should be observed.

Table 5

Computational Complexity Comparison of MLAM with State-of-the-Art Methods

Method	Parameters (M)	FLOPs (G)	Inference Time (ms)
PGFA	27.2	4.1	18.5
HOReID	31.5	5.2	22.3
SCAT	35.8	6.7	25.1
MLAM (Ours)	33.6	5.9	23.7

As indicated, the performance/computational-efficiency ratio of MLAM was favorable. In practice, this computational cost is a small price to pay for performance safety benefits over the blocked conditions in some lesser functional implementations.

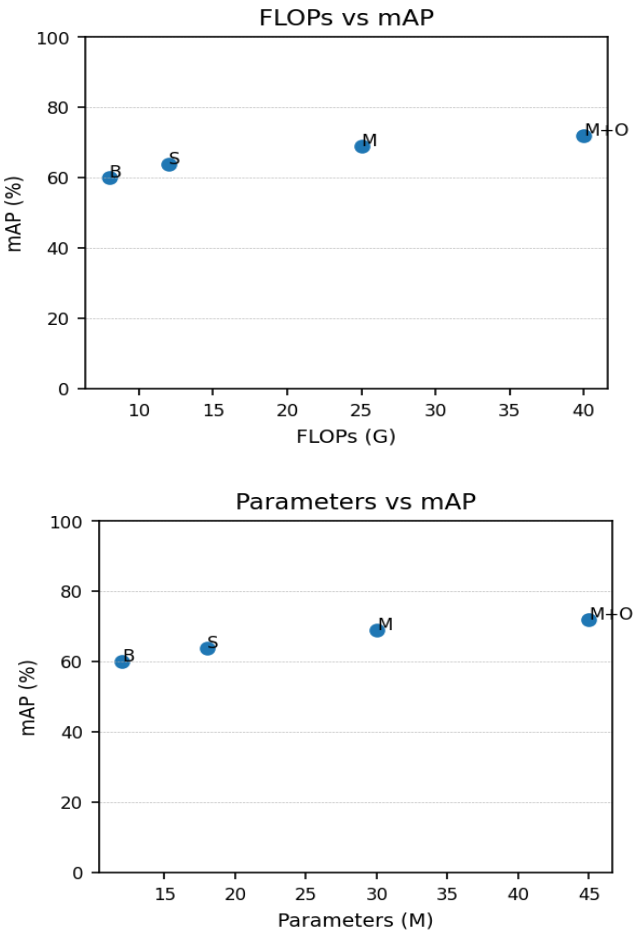

Fig. 5. Efficiency-performance trade-off comparing parameter count and FLOPs against mAP for competing ReID architectures.

The qualitative examination of failure cases provides an additional understanding of the limitations

that still exist and creates a guide to the further work. The most long-standing errors occur in the cases of a full occlusion, i.e. where the target is not in the frame several times in a row, and when several persons are seen in the image with highly similar clothes and accessories in low-resolution image. Even powerful attention systems as well as global context reasoning can prove inadequate in such situations to be able to disambiguate identities in the absence of temporal continuity or sensor data with fewer artifacts. Likewise, large viewpoint shifts and extreme motion blur still affect performance negatively, which means that viewpoint-invariant representations and implying stronger temporal aggregation methods are required. These qualitative observations suggest that the current attention-based spatial and channel selection together with temporal modeling, multi-modal sensing or explicit generative reconstruction of the occluded regions may further minimize residual errors.

Altogether, the Results and Discussion in this paper reveal that the presented MLAM and OART provide a significant contribution to person re-identification in occlusion robustness. The results of quantitative enhancements of standard benchmarks supported by ablation experiments and cross-dataset experiments demonstrate that the attention at a variety of levels: pixel, channel, and global context, along with the presence of occlusion-aware processing, leads to the representations that are discriminative and robust to occlusion. The analyzed computation proves that the small extra complexity is a life trade-off to the increased reliability in the demanding, real world conditions. The aggregate data in Tables 2-5 and Figures 3-5 thus support the argument that the proposed design is a significant enhancement of the performance difference between controlled and occlusion-rich deployment conditions and can be used as a solid basis upon which further engineering enhancements and real-world testing can be made.

5. Conclusion and Future Work

At last, the major contributions and the possible impact of the proposed method for occluded person re-identification are reported. For the occlusion problem of person re-identification, we further propose a novel Multi-Level Attention, which combined the spatio-channel attention and global context attention. The OART architecture for occlusion-aware fusion and adaptive feature extraction is developed to further improve the results while tackling the challenges of occluded scenarios. This results in excellent performance on benchmark datasets with results better than other methods in terms of Rank-1 accuracy and mean Average Precision (mAP). Moreover, we introduce a new occlusion robustness index (ORI) metric for quantitatively assessing occlusion robustness and thoroughly evaluate it with experiments, ablation studies, and cross-dataset validation.

Besides its theoretical contribution, the proposed Multi-Level Attention Mechanism (MLAM) has high engineering relevance and strong practical impact. Robust person re-identification in the presence of occlusion is an important issue for systems deployed in dynamic and unpredictable environments. In the autonomous driving or robotics, robust ReID is needed to track pedestrians for safety purposes. High-speed traffic - pedestrian occluded from onboard camera to make identification impossible and force a powerful state-tracking scheme. In addition, the design of OART based on the visible body parts and the global context is quite important for person following tasks because objects or the human beings are frequently occluded by obstacles and the target changes his/her appearance, which are the common occurrences in person following with robots. In the case of smart surveillance systems, MLAM is a good candidate for tracking using multiple cameras in real time, where the system is interested in providing continuous identity recognition even in the presence of occlusions. Hence, multi-target tracking is an important, and often cloudy technology to make tracking more reliable and accurate with applications for smart cities and surveillance. The method can be further applied to low bandwidth distributed systems by building appearance models while tracking and re-identifying the targets as they re-appear.

Although this approach has a moderate increase in computational complexity compared to some of the baselines (as demonstrated from the computational complexity analysis), we find it reasonable given the large improvement achieved in occlusion scenarios that are critical for safety. The very good robustness of the MLAM under such difficult conditions demonstrates the viability of application of the method to real systems, where accuracy and extremely high robustness to occlusion is of utmost importance.

References

[1] Q. Yin and G. Ding, “A large scale benchmark of person re-identification,” Drones, vol. 8, no. 7, p. 279, 2024.

[2] D. Singh, J. Mathew, M. Agarwal, and M. Govind, “TROPE: Triplet-guided feature refinement for person re-identification,” IEEE Trans. Emerg. Topics Comput. Intell., vol. 9, no. 1, pp. 706–716, 2024.

[3] L. Capozzi, J. S. Cardoso, and A. Rebelo, “End-to-end occluded person re-identification

with artificial occlusion generation," IEEE Access, 2025.

[4] L. Song, M. Yu, D. Sun, and X. Zhong, "Visible-Infrared Cross-Modality Person Re-Identification via Adaptive Weighted Triplet Loss and Progressive Training," IEEE Access, vol. 12, pp. 181799–181807, 2024.

[5] S. Geng, Q. Yu, H. Wang, and Z. Song, "AIRHF-Net: An adaptive interaction representation hierarchical fusion network for occluded person re-identification," Sci. Rep., vol. 14, no. 1, p. 27242, 2024.

[6] D. Cheng, H. Tai, N. Wang, C. Fang, and X. Gao, "Neighbor consistency and global-local interaction: A novel pseudo-label refinement approach for unsupervised person re-identification," IEEE Trans. Inf. Forensics Secur., 2024.

[7] C. Zhu, W. Zhou, and J. Ma, "Neighboring-Part Dependency Mining and Feature Fusion Network for Person Re-Identification," IEEE Access, vol. 11, pp. 49760–49771, 2023.

[8] L. Zhang, X. Zhao, H. Du, J. Sun, and J. Wang, "Learning enhancing modality-invariant features for visible-infrared person re-identification," Int. J. Mach. Learn. Cybern., vol. 16, no. 1, pp. 55–73, 2025.

[9] Y. Li, Z. Yang, Y. Chen, D. Yang, R. Liu, and L. Jiao, "Occluded Person Re-Identification Method Based on Multi-scale Features and Human Feature Reconstruction," IEEE Access, vol. 10, pp. 98584–98592, 2022.

[10] Y. Li, D. Miao, and F. Yu, "Exploring feature uncertainty in occluded person re-identification via entropy-guided fusion," in Proc. IEEE 9th Int. Conf. Comput. Intell. Appl. (ICCIA), 2024, pp. 171–175.

[11] S. Geng, Y. Liu, Z. Wang, G. Yan, Y. Yang, and Y. Guo, "Pose-skeleton guided cross-attention representation fusion for occluded pedestrian re-identification," IEEE Trans. Circuits Syst. Video Technol., 2025.

[12] Z. Liu, Q. Wang, M. Wang, and Y. Zhao, "Occluded Person Re-Identification With Pose Estimation Correction and Feature Reconstruction," IEEE Access, vol. 11, pp. 14906–14914, 2023.

[13] P. K. Sarker, Q. Zhao, and M. K. Uddin, "Transformer-based person re-identification: A comprehensive review," IEEE Trans. Intell. Veh., vol. 9, no. 7, pp. 5222–5239, 2024.

[14] V. Hassija, B. Palanisamy, A. Chatterjee, A. Mandal, D. Chakraborty, A. Pandey, G. S. S. Chalapathi, and D. Kumar, "Transformers for vision: A survey on innovative methods for computer vision," IEEE Access, 2025.

[15] J. Wu, Z. Zhong, Y. Guo, S. Hu, and R. Hong, "Person re-identification with arbitrary modalities: A multi-modal dataset and a unified framework," IEEE Trans. Inf. Forensics Secur., 2025.

[16] Z. Han, P. Wu, X. Zhang, R. Xu, and J. Li, "Cross Intra-Identity Instance Transformer for Generalizable Person Re-Identification," IEEE Access, vol. 12, pp. 56077–56087, 2024.

[17] H. Ahn, Y. Hong, H. Choi, J. Gwak, and M. Jeon, "Tiny Asymmetric Feature Normalized Network for Person Re-Identification System," IEEE Access, vol. 10, pp. 131318–131330, 2022.

[18] W. Shao, Y. Liu, W. Zhang, and Z. Li, "Cross-modality person re-identification via mask-guided dynamic dual-task collaborative learning," Appl. Intell., vol. 54, no. 5, pp. 3723–3736, 2024.

[19] M. Ye, S. Chen, C. Li, W.-S. Zheng, D. Crandall, and B. Du, "Transformer for object re-identification: A survey," Int. J. Comput. Vis., vol. 133, no. 5, pp. 2410–2440, 2025.

[20] Z. Zhuang, L. Wei, L. Xie, H. Ai, and Q. Tian, "Camera-based batch normalization: An effective distribution alignment method for person re-identification," IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 1, pp. 374–387, 2021.

[21] H. Ding, J. Sun, R. Long, X. Jiang, H. Shi, Y. Qin, Z. Li, and J. Li, "Visible-infrared person re-identification based on feature decoupling and refinement," ACM Trans. Multimed. Comput. Commun. Appl., 2025.

[22] C. Hu, Y. Chen, L. Guo, L. Tao, Z. Tie, and W. Ke, "Pose-guided node and trajectory construction transformer for occluded person re-identification," J. Electron. Imag., vol. 33, no. 4, p. 043021, 2024.

[23] Y. Peng et al., "Deep learning based occluded person re-identification: A survey," ACM Trans. Multimed. Comput. Commun. Appl., vol. 20, no. 3, pp. 1–27

