Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28502

Enhancing Global Features with Fine-Grained Local Features for Occluded Person Re-identification

Zixu Wang

School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
22281111@bjtu.edu.cn

Abstract. Occluded Person Re-identification (Occluded Re-ID) is a key challenge in video surveillance, where partial occlusions degrade matching accuracy by causing missing appearance cues. Existing methods often focus on reconstructing occluded regions, which may introduce artifacts and underuse reliable visible fine-grained features. To solve this, we propose a multi-level local-to-global enhancement framework. We simulate occlusions via random erasing/cropping, extract global and patch features using a Vision Transformer (ViT), refine patches into four local groups with a lightweight Local–Global Relation Module, and sequentially fuse refined locals into the global representation. Multi-branch supervision is applied to global and local features. Experiments on Occluded-DukeMTMC show our method outperforms state-of-the-art approaches, achieving 72.0% Rank-1 accuracy and 62.5% mAP, confirming its robustness to occlusions.

Keywords: Occluded Person Re-identification, Local Feature Enhancement, Global Feature Representation, Vision Transformer (ViT), Global–Local Fusion

1. Introduction

Occluded Person Re-Identification (Occluded Re-ID) is crucial for video surveillance and intelligent security, where a target must be matched across cameras despite partial visibility. In realistic scenes, static objects, other pedestrians, and scene layouts frequently occlude body parts, yielding missing or distorted appearance cues and causing a notable drop in matching robustness and accuracy. This motivates approaches that remain reliable when only a subset of the pedestrian is visible.

Most existing methods address occlusion by recovering features in the occluded regions, for example through mask-guided reconstruction, completion networks, or generative adversarial models [1,2]. Such strategies aim to compensate for the missing information, and in some cases are able to restore plausible global structures of pedestrian images. However, these methods often suffer from two fundamental limitations. First, they tend to under-exploit fine-grained features from visible regions, which are the most immediate, reliable, and noise-free cues for pedestrian matching. Second, the process of reconstructing occluded features is inherently uncertain; imperfect recovery may introduce artifacts or semantic noise, which can propagate into the global descriptor and mislead the matching process. As a result, while reconstruction-based approaches may improve recall under mild occlusions, their reliance on uncertain synthetic features limits robustness in more complex and heavily occluded real-world scenarios. This reveals the necessity of shifting focus

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28502

toward maximizing the discriminative potential of visible information rather than primarily relying on feature completion.

Based on the aforementioned concerns, we propose Enhancing Global Features with Fine-Grained Local Features for Occluded Person Re-identification, a multi-level local-to-global enhancement framework that explicitly amplifies the contribution of visible-region details in the final global descriptor. Concretely, we: (1) simulate occlusions via random erasing and random cropping at the data level to encourage robustness; (2) extract global and patch-level features with a ViT backbone, randomly partition the patch features into four local groups, and for each group apply a Local–Global Relation Module that correlates local features with the global feature (Conv1×1 projections + dot-product relation + sigmoid gating), producing a refined local feature and (3) sequentially fuse the four refined locals back into the global representation. We supervise the final global feature, its erasing/cropping counterparts, and all intermediate locals with cross-entropy and triplet losses to enforce discriminability at both global and local levels.

We summarize our contributions as follows:(1)A multi-level local feature enhancement paradigm that strengthens visible, fine-grained cues within the global descriptor under occlusion.(2)A lightweight Local–Global Relation Module (Conv1×1 projections + relation gating) and sequential fusion strategy that integrate local refinements into the final global feature.(3)A multi-branch supervision scheme that applies CE + Triplet to the final global, its random-erased/cropped variants, and all intermediate local features, improving robustness and generalization to occlusions.

2. Related work

2.1. Traditional person re-identification

Traditional Re-ID assumes mostly visible pedestrians and focuses on robust, discriminative representations invariant to viewpoint, illumination, and background changes. Early approaches used handcrafted descriptors with metric learning; modern methods rely on deep architectures to learn global descriptors and sometimes part-based features.

Representative methods can be broadly categorized into four directions. Early handcrafted approaches combined with metric learning, such as LOMO+XQDA [3], capture color and texture statistics and learn a cross-view metric to match pedestrians. With the rise of deep learning, global deep embedding methods such as IDE [4] treat Re-ID as a classification problem, using a softmax objective to learn ID-discriminative features, often enhanced with metric losses such as triplet loss. Part-based models, exemplified by PCB [5], horizontally partition the pedestrian image into several stripes to extract local features, which are then concatenated into a stronger global descriptor. More recently, attention-based mechanisms such as HA-CNN [6] and its variants refine the global representation by highlighting identity-relevant regions, thereby improving robustness against background noise and irrelevant information.Limitation for occlusion. When severe occlusion occurs, methods relying on holistic cues degrade because critical body parts are missing; naive global pooling also dilutes sparse visible evidence. This motivates explicit modeling of visibility and stronger utilization of local fine-grained features. In addition, works such as Part-Aligned bilinear representations and Part Bilinear pooling [7] further enhance local feature modeling. Random Erasing augmentation [8] has also been shown to improve robustness against partial occlusions.

DOI: 10.54254/2755-2721/2025.LD28502

2.2. Occluded person re-identification

Occluded Re-ID tackles partial visibility by either (a) localizing and down-weighting occlusions, (b) completing missing content, or (c) enhancing visible regions to dominate matching.

Representative methods for occluded person re-identification generally fall into four categories. The first line of research focuses on occlusion localization and masking, e.g., PGFA [9] and PVPM [10], where visibility or part masks are employed to guide feature extraction and suppress unreliable regions. Another direction is feature or image completion, e.g., FD-GAN [2] and Adver Occluded [1], which uses pose-guided alignment or GAN-based models to reconstruct missing body parts, though such reconstructions may introduce artifacts and noise. A third category emphasizes relation and part modeling, such as HOReID [11], in which approaches such as high-order relation modeling establish correspondences between visible body parts across views. Finally, attention-based visible enhancement methods apply region-level attention mechanisms to increase the contribution of visible cues within the global descriptor, e.g., RGA-SC [12], thereby improving robustness under occlusion.Gap addressed by our method. Prior works often over-rely on reconstruction or integrate local cues in a limited, single-scale manner. Our framework directly amplifies visible, fine-grained local features through explicit local–global relation gating and sequential fusion into the global descriptor, while data-level occlusion simulation (random erasing/cropping) plus multi-branch supervision further regularize the model to favor reliable visible evidence.

3. Method

3.1. Overview

Given an input pedestrian image $\mathrm{I}\in\mathrm{R}^{\mathrm{H}\times\mathrm{W}\times3}$, the goal of our method is to learn a discriminative feature representation that remains robust under occlusion. We employ a Vision Transformer (ViT) backbone $\mathrm{F}(\cdot)$ to extract both global features and patch-level features. Specifically

$$f_g = F_{global}(I),\ \{p_1, p_2, \ldots, p_N\} = F_{patch} \tag{1}$$

where $\mathrm{f_g}\in\mathrm{R}^{\mathrm{d}}$ is the global feature vector and $\mathrm{p_i}$ denotes the patch embeddings from ViT.

The patch embeddings are randomly partitioned into four local groups $\left\{\mathrm{f_l^{(1)}},\mathrm{f_l^{(2)}},\mathrm{f_l^{(3)}},\mathrm{f_l^{(4)}}\right\}$, each of which is processed by a Local Feature Enhancement Module (LFEM). The enhanced features are then sequentially integrated with the global representation to produce the final feature $\mathrm{f_{final}}$ .To improve robustness against occlusions, we additionally apply random erasing and random cropping augmentations on the input images, producing augmented views that are trained jointly.Finally, both the global and local features are optimized using a multi-loss scheme that combines cross-entropy and triplet loss, ensuring identity discriminability and metric separability.The overall framework of our proposed method is illustrated in Figure 1, which shows the input images with augmentations, the Vision Transformer [13] backbone, the Local Feature Enhancement Module (LFEM), the sequential fusion into the enhanced global descriptor, and the multi-branch loss supervision.

DOI: 10.54254/2755-2721/2025.LD28502

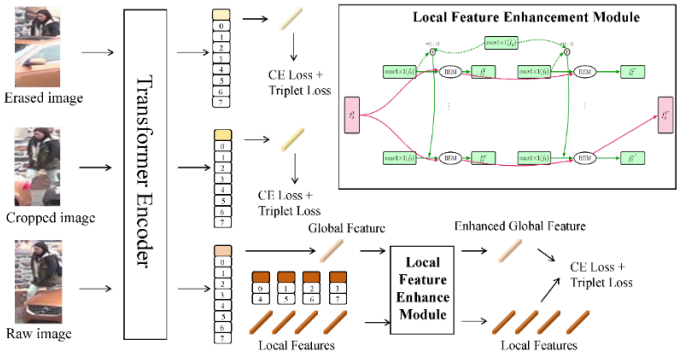

Figure 1. The overall framework

3.2. Data augmentation

Occlusion is highly variable in real-world surveillance, making it essential to simulate different occlusion patterns during training. We employ two augmentation strategies: Random Erasing (RE): With a certain probability, a rectangular region of the input image is randomly selected and replaced with random noise. This forces the model to rely on non-occluded parts for identity recognition. Random Cropping (RC): A random sub-region of the image is cropped and resized back to the original resolution, simulating partial visibility from bounding box misalignment or camera viewpoints.

Formally, for each input image I we generate:

$$I_{RE}, \quad I_{RC}$$

which are passed through the same feature extraction and training pipeline as the original I.

3.3. Local feature enhancement module

When occlusion occurs, relying solely on global features is unreliable, as important cues may be missing. To mitigate this, we design a Local Feature Enhancement Module (LFEM), which adaptively strengthens visible local features by modeling their relation with the global context.

Given a local feature f_l and the global feature f_g, the LFEM operates as follows:

$$f_{l1}=f_l W_{l1}, \quad f_{l2}=f_l W_{l2}, \quad f_g^{'}=f_g W_g$$

where W_{l1}, W_{l2}, W_g are learnable projection matrices.

We compute a relation weight by a dot product followed by a sigmoid activation:

$$r=\sigma\left(f_{l1}\cdot f_g^{'}\right)$$

The refined local contribution is then:

$$f^{'}=r\cdot\left(f_{l2}W_l^{'}\right)$$

DOI: 10.54254/2755-2721/2025.LD28502

where $W_l^{'}$ is another learnable projection.

Finally, the enhanced local-global fusion feature is obtained as:

$$f_{final}=f_l+f_g+f^{'}$$

Each of the four local groups $\left\{f_l^{(1)},f_l^{(2)},f_l^{(3)},f_l^{(4)}\right\}$ is processed by LFEM, and their outputs are integrated sequentially into the global feature, yielding a robust final descriptor.

3.4. Loss function

We adopt a multi-loss objective to optimize both classification discriminability and metric separability. Cross-Entropy Loss (L_{CE}) is applied to predict the pedestrian identity labels, guiding the model to learn features that are discriminative across different IDs.

$$L_{CE}=-\sum_{i=1}^{M} y_i \log \hat{y}_i$$

where y_i is the ground-truth label and $\hat{y}_i$ is the predicted probability. Triplet Loss (L_{Tri}) enforces a margin between positive and negative pairs, ensuring that features of the same identity are closer than those of different identities:

$$L_{Tri}=\sum\left[|f_a-f_p|_2^2-|f_a-f_n|_2^2+\alpha\right]_+$$

where f_a,f_p,f_n denote anchor, positive, and negative features, and α is the margin.

We apply both losses to: the final global feature f_{final} , augmented features f_{RE},f_{RC} and intermediate local features from LFEM

The overall loss function is:

$$L=\sum_{k\in\left\{final,RE,RC,l^{(1)},l^{(2)},l^{(3)},l^{(4)}\right\}}\left(L_{CE}^{(k)}+L_{Tri}^{(k)}\right)$$

This multi-branch supervision encourages discriminability at both local and global levels, leading to a representation that is robust to occlusion.

4. Experiments

4.1. Dataset and evaluation metric

We evaluate our method on the widely used Occluded-DukeMTMC (Occluded-Duke) dataset, which is specifically designed for occluded person re-identification. The dataset contains $15,618$ training images of 702 identities and $17,661$ test images of another 702 identities, where a large proportion of test samples are occluded.

For evaluation, we adopt Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) as the metrics. Specifically, the Rank- k accuracy measures whether the correct identity appears in the top-kk retrieved results:

$$\text{Rank-k}=\frac{1}{N}\sum_{i=1}^{N}1\left[\exists j\in\{1,\ldots,k\},\text{id}\left(g_j\right)=\text{id}\left(q_i\right)\right]$$

DOI: 10.54254/2755-2721/2025.LD28502

where q_i is the i-th query image, g_j is the j-th gallery result, and $1[\cdot]$ is the indicator function.

The mAP metric evaluates retrieval performance by averaging precision over all recall levels:

$$mAP=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{j=1}^{m_i}\text{Precision}(j)$$

where m_i is the number of relevant gallery images for query q_i.

4.2. Implementation details

We implement our method on top of the TransReID framework [14] with a Vision Transformer backbone. Specifically, we adopt ViT-Base Patch16 [15] (TransReID variant) pretrained on ImageNet as the backbone network. The backbone is configured with a stride size of $[11,11]$, and incorporates the camera-aware Side Information Embedding (SIE) module with coefficient 3.0. In addition, the Jigsaw Patch Module (JPM) is enabled together with a feature rearrangement mechanism.

The input images are resized to 256×128. For data augmentation, we employ random horizontal flipping with probability 0.5, random erasing with probability 0.5, and zero-padding of 10 pixels. During training, we use the PK-sampler, where each mini-batch contains 16 identities with 3 images per identity, resulting in 48 images per batch.

Optimization is performed using Stochastic Gradient Descent (SGD) with momentum. The initial learning rate is set to 8×10^{-3}, with weight decay of 1×10^{-4} and bias weight decay of 1×10^{-4}. The learning rate for bias parameters is further scaled by a factor of 2. A linear warm-up strategy is applied in the early epochs, followed by a cosine decay schedule. The model is trained for 200 epochs, with checkpoint saving and evaluation performed every 20 epochs.

The loss function is a combination of cross-entropy loss for identity classification and triplet loss with margin parameter $\alpha=0.3$. During testing, we extract features before the BN neck layer, followed by L2 normalization. Re-ranking is disabled. All experiments are conducted on a NVIDIA GeForce RTX 5090 GPU.

Table 1. Comparison with SOTA methods

Model	Rank-1	Rank-5	Rank-10	mAP
LOMO+XQDA	8.1	17.0	22.0	5.0
DIM	21.5	36.1	42.8	14.4
Part Aligned	28.8	44.6	51.0	20.2
Random Erasing	40.5	59.6	66.8	30.0
HACNN	34.4	51.9	59.4	26.0
PCB	42.6	57.1	62.9	33.7
Part Bilinear	36.9	-	-	-
Adver Occluded	44.5	-	-	32.2
FD-GAN	40.8	-	-	-
DSR	40.8	58.2	65.2	30.4
SFR	42.3	60.3	67.3	32.0
Ours	72.0	83.3	87.5	62.5

DOI: 10.54254/2755-2721/2025.LD28502

4.3. Comparison with SOTA methods

We compare our method with several state-of-the-art person re-identification approaches. Among them, LOMO+XQDA [3], DIM [16], Part Aligned [7], Random Erasing [8], HACNN [6], PCB [5], and Part Bilinear [7] are regarded as traditional Re-ID methods, as they were originally designed for the general setting without explicitly modeling occlusion. In contrast, Adver Occluded [1], FD-GAN [2], DSR [17], SFR [18] and our method represent occlusion-aware Re-ID methods, which explicitly address the challenges of partial visibility through reconstruction, masking, or attention mechanisms. Results are reported in terms of Rank-1, Rank-5, Rank-10 accuracy and mAP on the Occluded-Duke dataset. As shown in Table 1, our method significantly outperforms existing approaches, particularly in mAP and Rank-1, demonstrating the strong discriminative capability of enhancing global features with fine-grained local features.

4.4. Ablation study

To further investigate the effectiveness of each component, we conduct ablation studies on the Occluded-Duke dataset. Starting from the ViT baseline, we examine the contribution of Local Feature Enhancement Module (LFEM), Data Augmentation, and the full integration of our proposed modules.As shown in Table 2, the base model achieves 58.2 mAP and 69.3% Rank-1 accuracy. Removing or disabling specific modules results in performance fluctuations, while our full model achieves the best results, with 62.5 mAP and 72.0% Rank-1 accuracy. This confirms that our local feature enhancement and multi-branch supervision contribute significantly to the robustness against occlusion.

Table 2. Ablation studies

Model	Rank-1	Rank-5	Rank-10	mAP
Baseline(ViT)	69.3	81.4	85.7	58.2
LFEM only	71.1	82.4	86.3	59.3
RE&RC only	71.6	83.5	87.1	62.3
Full	72.0	83.3	87.5	62.5

5. Conclusion

In this paper, we presented a novel framework for Enhancing Global Features with Fine-Grained Local Features to address the challenging problem of occluded person re-identification. Unlike previous approaches that mainly focus on recovering occluded regions, our method emphasizes fully exploiting visible-region information through a multi-level local-to-global enhancement strategy. By integrating local features with global representations via the proposed Local Feature Enhancement Module, the model effectively strengthens the reliability of visible cues while maintaining global contextual awareness.

Furthermore, we incorporated occlusion-oriented data augmentation (random erasing and random cropping) and a multi-branch loss supervision scheme, ensuring discriminability at both local and global levels. Extensive experiments on the Occluded-Duke dataset demonstrated that our approach outperforms existing state-of-the-art methods, achieving significant improvements in both Rank-1 accuracy and mAP.

DOI: 10.54254/2755-2721/2025.LD28502

In the future, our framework can be extended to other vision tasks where partial visibility and feature incompleteness are critical issues, such as crowd analysis, human–object interaction recognition, and multi-camera tracking. We believe our work offers a promising step toward more robust and generalizable solutions for real-world surveillance applications.

References

[1] Zhuo J, Chen Z, Lai J, et al. Occluded person re-identification [C]. 2018 IEEE international conference on multimedia and expo (ICME), 2018: 1-6.
[2] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro [C]. Proceedings of the IEEE international conference on computer vision, 2017: 3754-3762.
[3] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 2197-2206.
[4] Xiao T, Li H, Ouyang W, et al. Learning deep feature representations with domain guided dropout for person re-identification [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 1249-1258.
[5] Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline) [C]. Proceedings of the European conference on computer vision (ECCV), 2018: 480-496.
[6] Li W, Zhu X, Gong S. Harmonious attention network for person re-identification [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 2285-2294.
[7] Suh Y, Wang J, Tang S, et al. Part-aligned bilinear representations for person re-identification [C]. Proceedings of the European conference on computer vision (ECCV), 2018: 402-419.
[8] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation [C]. Proceedings of the AAAI conference on artificial intelligence, 2020: 13001-13008.
[9] Miao J, Wu Y, Liu P, et al. Pose-guided feature alignment for occluded person re-identification [C]. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 542-551.
[10] Gao S, Wang J, Lu H, et al. Pose-guided visible part matching for occluded person reid [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 11744-11752.
[11] Wang G A, Yang S, Liu H, et al. High-order information matters: Learning relation and topology for occluded person re-identification [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 6449-6458.
[12] Zhang Z, Lan C, Zeng W, et al. Relation-aware global attention for person re-identification [C]. Proceedings of the ieee/cvf conference on computer vision and pattern recognition, 2020: 3186-3195.
[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.
[14] He S, Luo H, Wang P, et al. Transreid: Transformer-based object re-identification [C]. Proceedings of the IEEE/CVF international conference on computer vision, 2021: 15013-15022.
[15] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv: 2010.11929, 2020.
[16] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark [C]. Proceedings of the IEEE international conference on computer vision, 2015: 1116-1124.
[17] Iodice S, Mikolajczyk K. Partial person re-identification with alignment and hallucination [C]. Asian conference on computer vision, 2018: 101-116.
[18] He L, Liang J, Li H, et al. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7073-7082.