China Instrument and Control Society
Instrumentation
OPEN ACCESS
INSTRUMENTATION, Vol. 13, No. 1, March 2026 https://doi.org/10.15878/j.instr.202600316

Article

Visibility-Guided Dual-Branch Network for Occluded Person Re-Identification

Menghan An*, Yanyan Zhang, Yu Qin

School of Electronics Information Engineering, Nanjing University of Information Science & Technology, Nanjing 21004, China

* Corresponding author email: 202412180585@nuist.edu.cn

Abstract: Person re-identification (Re-ID) is a fundamental task in intelligent video surveillance, with widespread applications in urban security and intelligent transportation. However, in real-world scenarios, occlusions caused by pedestrians or environmental objects often degrade visual features, significantly impacting the robustness and accuracy of Re-ID systems. To address this challenge, we propose a Visibility-Guided Dual-Branch Network (VGDNet) for person occluded Re-ID. The network integrates a visibility-aware mechanism to dynamically guide feature extraction, adaptively balancing the reliance on global and local features under varying occlusion levels. Specifically, a global complementary branch is designed to capture holistic semantic information through explicit–implicit feature mining, while a dual-path local branch enhances fine-grained representations via localized reorganization and concatenation. To optimize feature discrimination, a multi-objective joint loss function is introduced by combining triplet loss, label-smoothed identity loss, and multi-similarity loss. Extensive experiments on three public benchmarks—Occluded-Duke, DukeMTMC-ReID, and Market-1501—demonstrate the effectiveness of the proposed method, achieving mAP/Rank-1 scores of 57.1%/64.3%, 83.1%/91.7%, and 91.0%/96.4%, respectively. Ablation studies and visualization further validate the contribution of each module in mitigating occlusion effects and enhancing feature robustness.

Keywords: person re-identification; occlusion; visibility guidance; dual-branch architecture

Citation: Menghan An, Yanyan Zhang, Yu Qin. "Visibility-Guided Dual-Branch Network for Occluded Person Re-Identification." Instrumentation 13, no.1 (March 2026). https://doi.org/10.15878/j.instr.202600316

1 Introduction

Person re-identification (Re-ID) [1-4] aims to accurately match image instances of the same individual captured from non-overlapping camera views. As a critical task in computer vision, Re-ID plays an essential role in intelligent surveillance, public safety, and smart city systems. The primary challenge in Re-ID lies in learning identity representations that are both highly discriminative and generalizable, enabling robust performance under complex real-world conditions, including cross-view appearance variations, partial occlusions, pose changes, and heterogeneous imaging environments.

In recent years, the rapid development of deep convolutional neural networks (CNNs) and vision Transformer architectures has significantly advanced the field of person Re-ID. Substantial improvements have been achieved in global feature representation, local region alignment, attention-based modeling, and structural information fusion. These advancements have notably enhanced the robustness and matching accuracy of Re-ID models. Nevertheless, under challenging conditions such as partial occlusion and severe pose variations, existing approaches still encounter difficulties in maintaining feature consistency and improving discriminative representation. These limitations underscore the necessity for more adaptive and structure-aware modeling strategies to ensure higher recognition

accuracy and stability in complex scenarios.

Occlusion constitutes a primary challenge in Re-ID, as it directly affects the completeness and reliability of feature extraction. As depicted in Fig. 1, pedestrians in real-world surveillance environments are frequently partially occluded by other individuals or surrounding objects, resulting in the loss or distortion of crucial identity cues. Such occlusions hinder global feature-based models from capturing comprehensive discriminative information, thereby impairing recognition performance. Additionally, the absence of visible body parts can lead to spatial misalignment in local feature-based methods, further degrading their effectiveness. In practical applications, various interrelated factors—including occlusion, pose deformation, viewpoint variation, illumination changes, background clutter, and low image resolution—collectively impose significant constraints on the accuracy and robustness of Re-ID systems. Addressing these challenges necessitates the development of robust, context-aware feature learning frameworks capable of coping with the inherent complexity of real-world conditions.

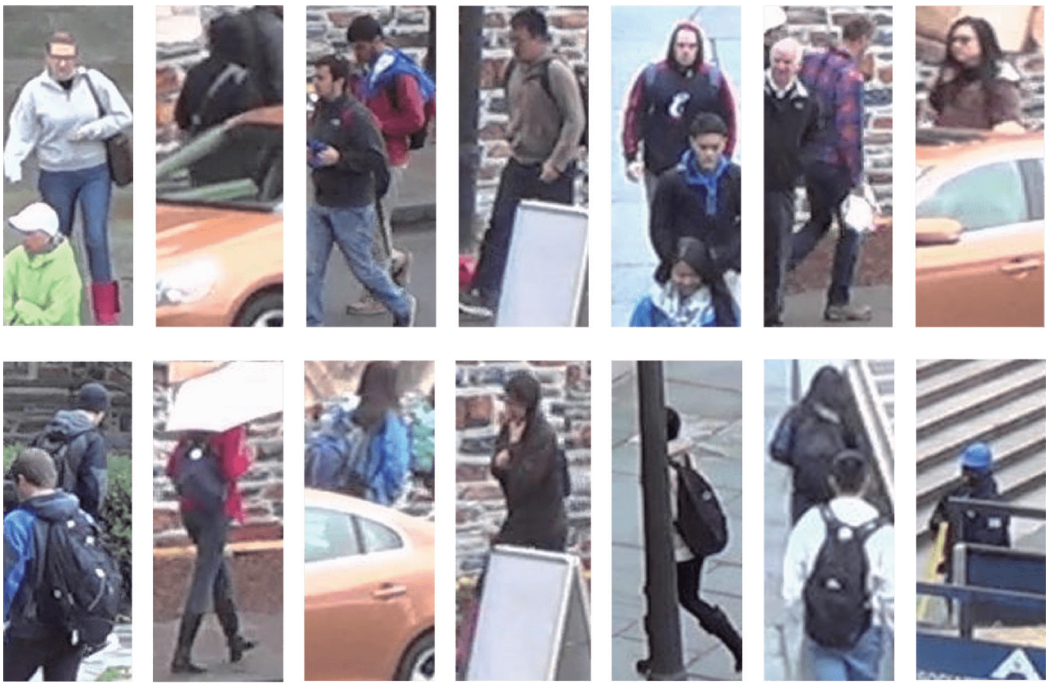

Fig.1 Example of occluded pedestrian images

To mitigate these limitations, researchers have proposed various strategies to enhance local feature modeling capabilities. Sun et al. [5] designed a partially convolutional baseline network that partitions pedestrian images into six horizontal regions for local feature extraction, thereby improving the model's ability to capture fine-grained details. However, this fixed horizontal partitioning methodology exhibits limitations in handling pose variations (e. g., arm movements), potentially assigning semantically consistent physical parts to different regions, thus compromising spatial partitioning consistency. To exploit multi-scale local information further, Fu et al. [6] proposed a horizontal pyramid matching framework that enhances local feature representation by fusing features across multiple horizontal partitioning scales. Although this approach enhances the model's discriminative capability, it remains vulnerable to noise interference in occluded regions, which ultimately degrades overall recognition performance. Conversely, Luo et al. [7] introduced an AlignedReID++ network, which employs a shortest path matching algorithm to achieve precise local feature alignment, mitigating error accumulation inherent in hard partitioning strategies. Nevertheless, its robustness under severe occlusion scenarios requires enhancement.

Current mainstream approaches exhibit limitations in addressing occlusion challenges in person Re-ID. Global feature-based representations are prone to background interference and lack robustness to occlusion due to insufficient modeling of local details, which hinders the extraction of critical identity information and consequently reduces recognition accuracy and overall robustness. While local feature-based representations enhance fine-grained perception and discriminative capability, they remain vulnerable to feature misalignment under occlusion or viewpoint variations, thereby constraining further model performance gains. Consequently, developing robust structural perception mechanisms with fine-grained local modeling capabilities constitutes a critical research objective for mitigating occlusion-induced feature incompleteness and information interference in person Re-ID systems.

To address these challenges, this work propose-s a Visibility-Guided Dual-Branch Network (VGDNet) for

occluded person Re-ID, designed to enhance discriminative capability and robustness in occlusion scenarios. The primary contributions comprise:

1. Integration of a position attention mechanism within the OSNet backbone, coupled with an attention entropy-derived visibility factor that quantifies occlusion severity in pedestrian images. This factor adaptively modulates feature weighting between global and local branches, enhancing feature selection precision and discriminative power under occlusion.

2. Within the dual-branch architecture, the global complementary branch enhances holistic semantic modeling by capturing contextual features inaccessible through local representations, improving overall pedestrian representation expressiveness and robustness. The local branch performs spatial reorganization and part-wise feature concatenation to capture structural relationships and fine-grained discriminative cues within salient regions, thereby enhancing the discrimination of local details.

3. To optimize feature discriminability and generalization, we propose a multi-objective joint loss function with dynamic weight allocation. This mechanism concurrently constrains and optimizes auxiliary objectives, facilitating task-aware feature disentanglement within the embedding space.

The proposed VGDNet for occluded person Re-ID incorporates a visibility-aware gating mechanism to modulate feature extraction pathways, enabling adaptive suppression of occluded regions and optimization of discriminative regions. This architecture demonstrates substantial performance advancements in occluded Re-ID scenarios. The VGDNet establishes a discriminative dual-branch framework with explicit structural awareness, while exhibiting significant scalability and deployment potential in real-world environments. Comprehensive experiments across multiple benchmarks confirm the method surpasses most state-of-the-art approaches in recognition accuracy for both occluded and holistic pedestrian images.

2 Related Work

2.1 Occluded Person Re-Identification

Occluded Person Re-identification is an important research branch of Re-ID, which specializes in solving the cross-camera matching challenges where pedestrian images are partially visible due to occlusion. Traditional Re-ID systems perform well in matching complete pedestrian images in cross-camera views, but in real surveillance scenarios, the prevalence of occlusion phenomena (e. g., object occlusion or crowd occlusion) causes the missing information of key body parts (e. g., head, torso, etc.), which seriously affects the completeness of the feature representations and the discriminative ability. The lack of local information caused by occlusion not only weakens the model's ability to discriminate target pedestrians, but also significantly increases the error rate of cross-view matching. Therefore, how to introduce stronger occlusion robustness into the model structure design, enhance the ability to focus on the visible region, and improve the feature extraction and discrimination performance in complex environments has become one of the core problems in the field of Re-ID.

To address the challenges of feature loss and degraded matching performance caused by occlusion, researchers both in China and abroad have conducted systematic studies focusing on key directions such as occlusion-aware enhancement and generation mechanisms, as well as pose estimation, and proposed a series of targeted improvement methods. The related methods are mainly introduced as follows:

One category of approaches enhances the representation of occluded images and improves the robustness of Re-ID systems under occlusion by leveraging occlusion-aware augmentation strategies and generative mechanisms. Wang et al.[8] proposed a Feature Erasing and Diffusion Network (FED), which employs an occlusion augmentation strategy to generate occlusion masks that guide the Occlusion Erasing Module (OEM) in removing non-pedestrian occlusion noise. In parallel, a Feature Diffusion Module (FDM) simulates multi-pedestrian interference in the feature space, thereby improving the model's perception of target pedestrians and its robustness against occlusion disturbances. Zhao et al. [9] proposed an Incremental Generative Occlusion Adversarial Suppression Network (IGOAS), which enhances the robustness to occlusion interference and the discriminative feature extraction for non-obscured areas of pedestrians. Qian et al.[10] proposed Pose Normalized Generative Adversarial Network (PN-GAN), which improves the robustness of the model to changes in pedestrian pose by generating pedestrian images with different poses, which can learn a new type of deep features without being affected by pose changes. Zhang et al. [11] proposed a Feature Completion Network (FeatComp), which automatically locates the occluded region through feature correlation modelling, guides the generator to perform feature completion on the invisible region, and optimizes the discriminator through adversarial training so that the completed features are difficult to differentiate from the real visible features, thus significantly reducing the intra-class variation due to occlusion.

Generative model-based methods have made some progress in improving the semantic integrity of occluded images, but there are still several challenges: for example, the training process is prone to instability, the semantic consistency of the generated images is insufficient, the network structure is more complex, and there is a stronger dependence on external a priori information. In addition, some of the methods focus too much on the

restoration effect at the image level and ignore the consistency modelling of local discriminative features, thus limiting their generalization ability and practical application effect under complex occlusion conditions.

Another category of approaches improves the recognition performance in occlusion scenarios by introducing a human pose estimation model to assist feature extraction and alignment. For example, Miao et al. [12] proposed a Pose-Guided Feature Alignment (PGFA) method, which enhances the robustness to occluded regions and the discriminative ability to non-occluded regions by guiding the local feature alignment through the human posture estimation, while Gao et al. [13] proposed a Pose-Guided Visible Part Matching (PVPM) method, which utilizes pose information to guide the attention mechanism to learn discriminative feature representations of individual body parts. Somers et al. [14] proposed a dual-supervised learning framework (BPBreID) that combines ID labels and human parsing labels to strengthen the discriminative nature of pedestrian features while using human parsing labels to guide the model to accurately locate the body parts, thus achieving a more robust pedestrian re-recognition in occlusion scenarios, but the human parsing labels also need to be introduced into the external pose estimation model, which adds an extra computational overhead.

Although existing methods improve the recognition performance in occluded scenes to some extent and enhance the model's ability to perceive the visible region, most of them rely on external auxiliary modules, such as human pose estimation or semantic segmentation networks. This dependency not only increases the structural complexity and computational resource consumption of the model, but also introduces additional training and inference overheads. In addition, there are usually significant domain differences between the training data relied on by the auxiliary modules and the real occlusion scenarios, which can easily lead to the degradation of the model's performance in real-world deployments, thus limiting the scalability and usefulness of this class of methods for occluded person Re-ID tasks.

2.2 Attention Mechanisms

The attention mechanism, as an important module in deep neural networks, can capture the key feature information of the target and enhance the model's ability to perceive discriminative regions. Fu et al.[15] proposed a dual attention network that fuses spatial and channel attention mechanisms to enhance feature representation in scene segmentation tasks. To better model global structural relationships, Zhang et al. [16] proposed a Relation-Aware Global Attention (RGA) module, which mines the global clustering structure by modelling bidirectional relationships between features, thus guiding the attention mechanism to focus on discriminative regions, but this method has high computational and memory overheads. To address the problem of pose change and image alignment, Li et al. [17] proposed Harmonious Attention CNN (HA-CNN) to jointly learn 'soft' pixel-level attention and 'hard' region-level attention to capture both global and local information. Fan et al. [18] proposed a Dual Structural Feature Network (DSF-Net), which dynamically discards the highest-attention region in the Euclidean structured branch through a feature discarding mechanism based on positional attention, thus directing the network to focus on other key regions. The network is guided to pay attention to other critical regions, thus avoiding the domination of feature learning by the occluded regions, and improving the performance of Re-ID in occluded scenarios.

The aforementioned methods improve the capability of attention mechanisms to model discriminative regions to some extent. However, in occluded scenarios, the occluded areas often occupy a larger portion of the image or exhibit strong interference, causing the network to allocate insufficient focus to the critical visible regions. This limitation ultimately degrades overall recognition performance. Consequently, enhancing the network's ability to emphasize key visible parts remains a crucial challenge in occluded person Re-ID.

Position Attention Mechanism (PAM) is able to enhance the correlation between pixels and improve the representation of high-attention regions by modelling the long-range dependencies between different spatial locations in an image. Compared with complex global modelling methods, PAM has a lightweight structure and low parameter overhead, which can enhance the perception of critical regions while maintaining computational efficiency. For this reason, this paper introduces the PAM module and designs the visibility factor based on it to evaluate the concentration degree of the attention distribution, in order to enhance the model's adaptability to the visibility change of the occluded region, and to improve the model's robustness in recognizing the human body region with the addition of a smaller number of parameters.

2.3 Loss Functions

In the Re-ID task, the loss function, as a metric mechanism to measure the difference between the model predictions and the real labels, plays a central role in guiding the optimization of the model parameters, improving the feature discrimination ability, and guaranteeing the training stability. Especially under the conditions of complex scenes such as occlusion, viewpoint change and background interference, the loss function is particularly crucial in strengthening model robustness. Therefore, the selection and design of the loss function becomes one of the key factors affecting the performance of the Re-ID system. In the occluded person Re-ID framework proposed in this paper, three types of loss functions with complementary properties are incorporated: label-smoothed identity loss, multi-similarity loss, and triplet loss. The design principles and

optimization mechanisms of the three types of loss functions are described in detail in the following section.

(1)Identity Loss for Label Smoothing

In the Re-ID task, Identity Loss (ID Loss) [19] transforms the Re-ID problem into a multi-classification problem by supervising the model to learn the features of different pedestrians through Cross-Entropy Loss. The approach essentially expects the model to accurately predict the corresponding identity label for each input image. Conventional cross-entropy loss uses a 'Hard Label', which assigns a value of 1 to the correct category and 0 to the rest, and the function is expressed as:

$$L_{ID} = -\sum_{i=1}^{N} y_i \log(p_i) \quad \begin{cases} y_i = 0, y \neq i \\ y_i = 1, y = i \end{cases} \tag{1}$$

where $i \in \{1, 2, ..., N\}$ denotes the pedestrian category, N denotes the number of pedestrian images in the training set, y is the true label, p_i is the predictive logic value of the network, and p_i can be expressed as:

$$p_i = \arg\max \frac{\exp\left(\left(W_{y_i}\right)^T x_i\right)}{\sum_{j=1}^{C} \exp\left(\left(W_j\right)^T x_i\right)} \tag{2}$$

Where x_i is the ith sample feature, C is the total number of label types, and W_j is a classifier for whether feature x_i belongs to label j, consisting of a fully connected layer. W_{yi} represents the classifier weight vector corresponding to the ground-truth label of the ith sample.

However, the traditional cross-entropy loss may cause the model to be overconfident in the prediction results at the late stage of training, resulting in overfitting and reducing the generalization ability of the model. By introducing Label Smoothing Regularization (LSR) to smooth the true labels and introduce appropriate uncertainty, the robustness and generalization ability of the model can be improved.

$$y_i = \begin{cases} y_i = 1 - \dfrac{N-1}{N}\varepsilon & if\ i = y \\ \varepsilon / N & otherwise \end{cases} \tag{3}$$

where ε is a hyperparameter with a small value that prevents the model from being overconfident in the training labels and reduces overfitting.

(2)Multi-similarity Loss

Multi-similarity Loss (MS Loss)[20] is a loss function for Deep Metric Learning (DML), which dynamically adjusts the weights of difficult samples to learn discriminative features more efficiently by simultaneously considering the three similarity relations (self-similarity, positive-sample relative similarity and negative-sample relative similarity) of the sample pairs. The formula for multiple similarity loss is as follows:

$$L_{MS} = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{\alpha}\log\left(1 + \sum_{j \in P_i} e^{-\alpha(S_{ij}-\lambda)}\right) + \frac{1}{\beta}\log\left(1 + \sum_{k \in N_i} e^{\beta(S_{ik}-\lambda)}\right)\right] \tag{4}$$

where m is the number of samples, S_{ij} is the cosine similarity between the anchor sample i and the positive sample j, S_{ik} is the cosine similarity between the anchor sample i and the negative sample k, P_i and N_i represent the sets of positive and negative samples of i, respectively. α, β are hyperparameters that control the weights of hard samples and λ is the similarity offset.

(3)Triplet Loss

Triplet Loss[21] is one of the most classical loss functions in deep metric learning, which makes samples of the same class closer together and samples of different classes further away in the feature space by learning the relative distance relationship between samples. Each triplet consists of three components: an anchor, a positive sample and a negative sample. The goal of the triplet loss is to ensure that the distance between the anchor and the positive sample is smaller than the distance between the anchor and the negative sample, which is usually controlled by a hyperparameter. The formula for triplet loss is expressed as follows:

$$L_{triplet} = \max\left(d(a,p) - d(a,n) + margin,\ 0\right) \tag{5}$$

where a is an anchor sample, p is a positive sample from the same class as a, n is a negative sample from a different class, and *margin* is a preset boundary hyperparameter controlling the minimum spacing between pairs of positive and negative samples. The goal of Triplet Loss is to make the distance $d(a,p)$ between a and p as small as possible and the distance $d(a,n)$ between a and n as large as possible.

3 The Proposed Model

To address structural information loss, feature degradation, and discriminative capability impairment induced by occlusion in person Re-ID, we propose VGDNet, which enhances discriminative robustness under complex occlusion scenarios through adaptive feature pathway modulation. The framework of VGDNet implements a visibility-aware mechanism that dynamically regulates feature extraction pathways according to varying occlusion patterns, enabling context-dependent weighting of global-local feature dependencies. The following sections detail the network architecture and submodule design specifications.

3.1 Overall Network Structure

The VGDNet architecture is illustrated in Fig.2. The framework is based on the improved OSNet backbone[22], leveraging its lightweight feature extraction structure. The input images undergo forward propagation through the OSNet backbone, with feature extraction terminating at the first convolutional layer (conv3_0) of Block 3. This design achieves model compression while preserving the low-level and mid-level semantic information expressiveness.

Following the backbone network conv3_0 layer, the architecture integrates a PAM with dual-branch feature extraction, and introduces a multi-objective joint loss

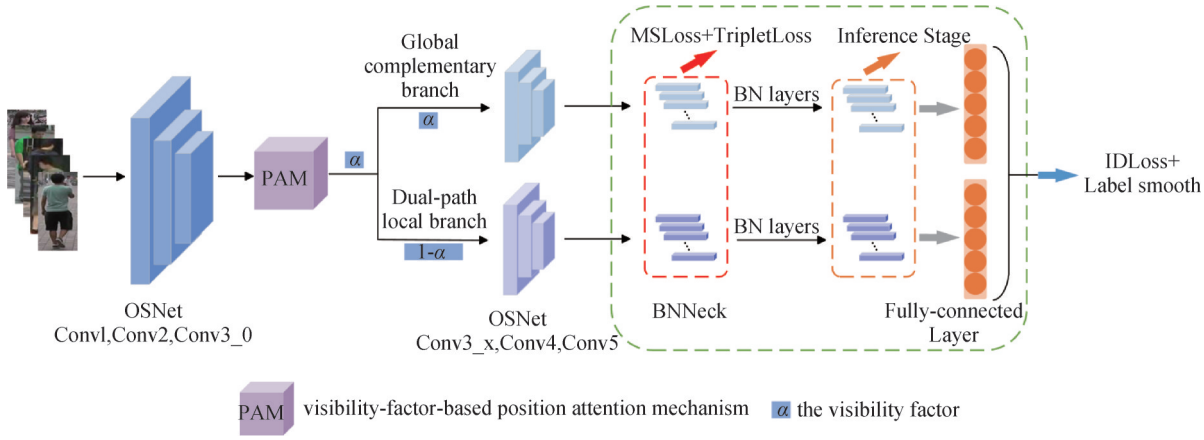

Fig.2 Overall network structure of the VGDNet

function for end-to-end optimization guided by the multi-granularity supervision mechanism. To enhance target region perception under occlusion, the PAM module captures global spatial dependencies within backbone features, emphasizing semantically salient regions to improve feature consistency and discriminative capacity. The enhanced features of the PAM module are processed by the two branches of global complementary and dual-path local in parallel, both inheriting OSNet's layer structure from conv3_0 through Block 5. This configuration generates unified feature representations.

Furthermore, a visibility factor (α) is formulated within the PAM module based on the normalized information entropy of the positional attention graph to quantify occlusion intensity in local feature map regions. This factor simultaneously characterizes the concentration of attention distribution and serves as a pivotal guidance signal to dynamically modulate collaborative modeling between dual-branch feature pathways. By leveraging the visibility factor to adjust inter-branch weighting coefficients, the model adaptively regulates the interdependence ratio of global and local features according to image-specific occlusion variations, thereby enhancing recognition stability and generalization capability under diverse occlusion scenarios. The global complementary branch augments holistic pedestrian representation modeling through an explicit-implicit dual-path mechanism: The explicit pathway preserves global structural information, while the implicit pathway employs Top DropBlock[23] to selectively suppress highly activated regions, compelling the network to discover latent features in low-response areas. The dual-path local branch enhances fine-grained feature extraction via a reorganization-concatenation mechanism, wherein the reorganization pathway utilizes visible regions to generate discriminative local representations, and the concatenation pathway reinforces semantic interdependencies between spatially contiguous regions to improve feature consistency under occlusion.

In addition, to achieve simultaneous optimization of pedestrian feature discriminability, robustness, and generalization capability, a multi-objective joint loss function is formulated. Under the multi-granularity supervision, MS Loss and Triplet Loss jointly optimize pre-BN layer[19] features within each branch, incorporating a learned weighting scheme to enhance intra-class compactness and inter-class separability in the metric space. Post-BN layer features are retained for inference to maintain the representation consistency. Features following the fully-connected layer are further subjected to label-smoothed ID Loss optimization to mitigate overfitting risk and strengthen generalization performance.

3.2 Visibility-Factor-based Position Attention Mechanism

A characteristic challenge in occluded pedestrian images is the spatially disorganized distribution of attention. When critical human body parts are obscured, the model lacks stable semantic cues, leading the attention mechanism to disperse focus more uniformly across the image. This behavior results in an increased entropy value within the attention map. By computing the average normalized information entropy of the entire feature representation, we can quantitatively assess the structural clarity of the image and the degree of occlusion. This metric effectively captures the image's interpretability and discriminative capacity in the feature space, thereby providing a global-level prior for occlusion-aware feature extraction.

A characteristic challenge in occluded pedestrian images is the spatially disorganized distribution of attention. When critical human body parts are obscured, the model lacks stable semantic cues, leading the attention mechanism to disperse focus more uniformly across the image. This behavior results in an increased entropy value within the attention map. By computing the average normalized information entropy of the entire feature representation, we can quantitatively assess the structural clarity of the image and the degree of occlusion. This metric effectively captures the image's interpretability and discriminative capacity in the feature space, thereby providing a global-level prior for

occlusion-aware feature extraction.

Leveraging the visibility factor as a dynamic modulation signal, the model adaptively regulates its reliance on global and local feature branches during training. This optimization adjusts feature extraction strategies according to occlusion conditions, mitigating performance degradation induced by occlusions. The visibility factor is formulated to address feature extraction impairment resulting from visibility uncertainty in target regions during occluded Re-ID tasks. It establishes a global regulation mechanism that dynamically senses occlusion intensity and modulates feature extraction pathway selection.

3.2.1 Position Attention Mechanism

The network architecture of the position attention mechanism based on the visibility factor is depicted in Fig. 3. Let the feature map output from the truncated OSNet backbone up to the conv3_0 convolutional layer be denoted as $F \in R^{C\times H\times W}$, where C represents the number of channels, and H and W denote the height and width of the feature map, respectively. The feature map F is subsequently processed through three separate 1×1 convolutional layers to produce three transformed feature representations: $X \in R^{C\times H\times W}$, $Y \in R^{C\times H\times W}$, and $Z \in R^{C\times H\times W}$, where C' is the reduced number of channels, typically set to C/8 to alleviate computational overhead.

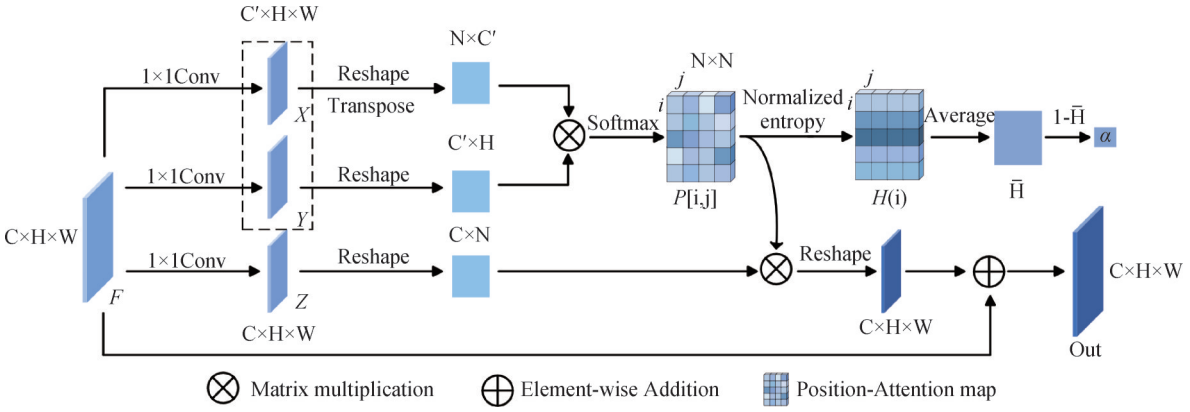

Fig.3 Network architecture of the visibility-factor-based position attention mechanism

Subsequently, the feature representations X and Y are reshaped to dimensions $X \in R^{C'\times N}$ and $Y \in R^{C'\times N}$, respectively, where $N = H \times W$. To capture the pairwise positional relationships across all spatial locations, X is transposed to $R^{N\times C'}$ and matrix-multiplied with Y. Finally, the positional attention weight matrix $P \in R^{N\times N}$ is obtained using the softmax layer. This process can be formally expressed as:

$$P_{[i,j]} = \frac{\exp(X_i \cdot Y_j)}{\sum_{j=1}^{N} \exp(X_i \cdot Y_j)},\ i,j \in \{1,\ldots,N\} \tag{6}$$

$P_{[i,j]}$ represents the attentional weight of the ith position to the jth position in the space. This attention weight matrix effectively captures the global spatial dependencies, thereby enhancing the representation of salient regions during feature modeling while suppressing the interference of irrelevant or redundant information.

After embedding the positional attention weights by multiplying the feature map Z with the attention matrix P, the result is superimposed with the original feature map to produce the final attention-enhanced feature representation Out, as formulated in Equation (7).

$$Out = \lambda Z \cdot P + F \tag{7}$$

Here, λ is a learnable weight parameter for attention modulation, and the output feature map Out is subsequently fed into the dual-branch network for further processing.

3.2.2 Construction of the Visibility Factor

As discussed in Section 2.1, the occluded Re-ID task poses a significant challenge wherein the network exhibits a tendency to allocate excessive attention to heavily occluded regions, consequently neglecting critical visible body parts and impairing overall recognition performance. Occluded regions are frequently associated with anomalous attention responses, characterized by the misallocation of attention toward irrelevant background or distractive elements. This spatial misalignment undermines the model's capacity to accurately localize semantically informative regions of the target. To address this limitation and enhance the model's occlusion-awareness, this work introduces an information entropy-based metric derived from the attention weights $P_{[i,j]}$ described in Section 3.2.1. Information entropy is employed to quantify the degree of disorder within the attention distribution, serving as a statistical measure of attention dispersion and uncertainty. Based on this formulation, a visibility scoring mechanism is established to evaluate the structural coherence and occlusion severity of input images, thereby providing a principled indicator to guide feature extraction under occlusion conditions.

The normalized information entropy is employed as a quantitative indicator of attention clutter at each spatial position. For the ith position, the entropy of the attention distribution is defined as:

$$H(i)=-\frac{1}{\log N}\sum_{j=1}^{N}P_{[i,j]}\times\log(P_{[i,j]}) \tag{8}$$

here *H(i)* denotes the attention entropy of the ith position, value $\in [0,1]$.

To evaluate the overall visibility of an image, the normalized entropy values across all spatial positions are averaged to compute the mean information entropy of the entire feature map:

$$\bar{H}=\frac{1}{N}\sum_{i=1}^{N}H(i) \tag{9}$$

Based on the quantitative analysis of the attention distribution's information entropy, the visibility factor $\alpha \in [0,1]$ is further defined as an inverse function of the attention entropy, serving as a complementary measure of the clarity and structural integrity of the feature map:

$$\alpha=1-\bar{H} \tag{10}$$

During the training phase, the visibility factor is utilized to adaptively modulate the model's dependence on the global complementary branch and the dual-path local branch. Specifically, the outputs from each branch are dynamically weighted through a visibility-aware modulation mechanism, thereby enabling the model to adjust feature aggregation strategies in accordance with occlusion conditions. The branch outputs under this adaptive weighting scheme are formally defined as follows:

$$\begin{aligned} glo' &= \alpha\cdot glo \\ par' &= (1-\alpha)\cdot par \end{aligned} \tag{11}$$

In which *glo* denotes the global complementary branch, which is designed to enhance the representation of global semantic information for occluded pedestrians via an explicit – implicit dual-path mining strategy. This branch aims to compensate for the semantic degradation introduced by occlusion through comprehensive global context modeling. In parallel, *par* represents the dual-path local branch, which focuses on extracting fine-grained discriminative features from localized regions. This design effectively alleviates the adverse impact of missing or attenuated local information under occlusion conditions, thereby preserving the model's discriminative capacity.

An illustrative example of information entropy computed from positional attention weights is depicted in Fig. 4. The positional attention matrix, with a dimensionality of 4×4, represents the spatial attention distribution of each location relative to all other positions within the feature map. Shannon entropy is employed to quantify the uncertainty associated with each row of the matrix, revealing significant variability in entropy values across different spatial positions. This variation reflects differing degrees of confidence in the attention distributions. Detailed interpretations are provided as follows:

The attention distributions corresponding to the first

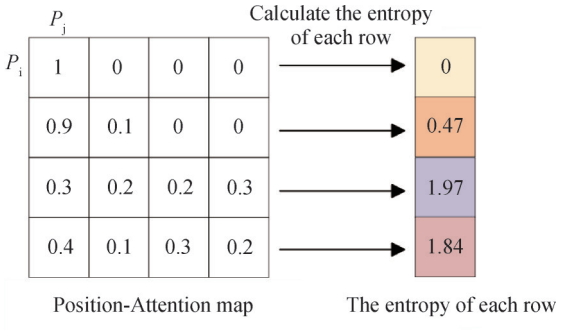

Fig.4 Example of information entropy with position attention weights

and second rows exhibit relatively high concentration, reflected by lower entropy values of 0 and 0.47, respectively. In particular, the first row assigns full attention to its own spatial position (with a self-weight of 1 and zero weights for all others), resulting in an entropy of 0. This indicates a completely deterministic attention distribution, signifying the model's high confidence in this spatial location. In contrast, the third and fourth rows demonstrate more dispersed attention patterns, with corresponding entropy values of 1.97 and 1.84. Notably, the attention weights in the third row approximate a uniform distribution ([0.3, 0.2, 0.2, 0.3]), yielding the highest entropy among all rows. This reflects the model's maximum uncertainty in attention allocation for that spatial position.

Overall, entropy values serve as effective quantitative indicators for characterizing both the concentration and uncertainty embedded in attention distributions, thereby facilitating the evaluation of image saliency and occlusion severity. As a statistical metric of distributional uncertainty, information entropy enables the assessment of attention focus, indirectly reflecting the model's ability to localize and exploit discriminative regions. Statistically, higher entropy reflects a more dispersed attention distribution, often caused by occlusions, background clutter, or weak discriminative cues. While lower entropy indicates a more focused attention, suggesting that the model is effectively focusing on salient and discriminative regions.

Consequently, based on the inverse quantization of $\bar{H}$, the visibility factor α exhibits a monotonic relationship with image occlusion: higher α values correspond to lower occlusion levels, while lower α values indicate more severe occlusion. When$\alpha\to 1$, the attention distribution becomes highly concentrated across spatial locations, with minimal occlusion and a distinctly visible target. Under these conditions, the model is inclined to prioritize global contextual features to facilitate identity recognition. Conversely, when$\alpha\to 0$, the attention distribution becomes increasingly dispersed, indicating the presence of occlusion or background interference and the lack of a clearly defined region of interest. Under such circumstances, the discriminative power of global features is substantially reduced, prompting the model to

place greater emphasis on local feature representations to ensure accurate identity recognition.

In contrast to existing re-identification approaches that rely on explicit occlusion annotations, pose estimation, or semantic segmentation modules, the proposed visibility factor construction method is designed to operate independently of external priors or additional supervisory signals, thereby offering enhanced adaptability and generalizability. Furthermore, the visibility factor serves as a modulation signal that guides the dual-branch architecture, enabling dynamic adjustment of the reliance between global and local feature branches. This facilitates occlusion-aware optimization in the feature extraction process, thereby improving the model's robustness under varying visibility conditions.

3.3 Global Complementary Branch

To address the limitations of Re-ID models in capturing implicit global features and their diminished robustness under complex occlusion conditions, a global complementary branch is introduced. This branch employs an explicit – implicit dual-path mining mechanism to systematically enhance the representation of global semantic information associated with occluded pedestrians. By reinforcing global contextual understanding, the proposed architecture significantly improves the model's discriminative capability in occlusion scenarios.

The global complementary branch structure, as illustrated in Fig. 5, processes the initial feature map of dimensions 512×24×8 obtained after the conv5 layer. It first performs explicit feature extraction using Generalized Mean Pooling (GeM). GeM employs a nonlinear spatial information aggregation strategy, preserving the most semantically discriminative regional features, thereby providing a clear and stable global representation $G \in R^{512\times1\times1}$ for the Re-ID task.

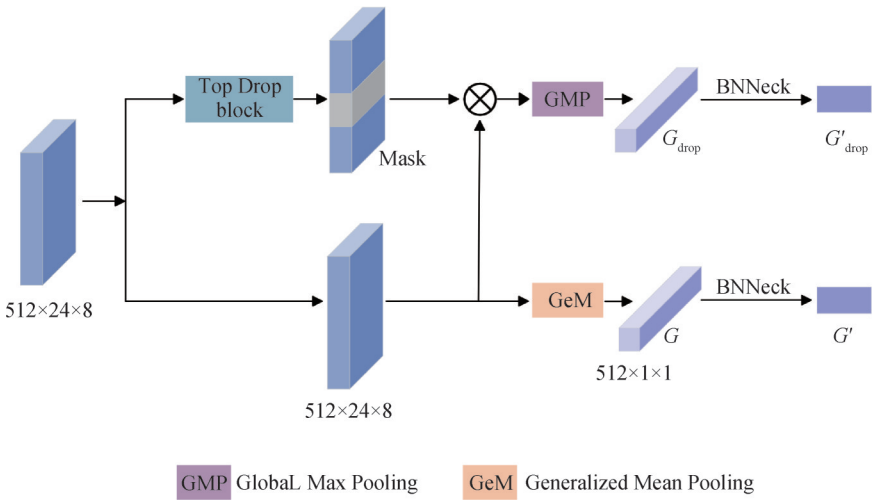

Fig.5 Network structure of the global complementary branch

To further exploit structurally informative yet weakly activated features within the spatial representation, an implicit modeling path is incorporated employing the Top DropBlock strategy. This path processes the initial feature map through a dropout module that selectively identifies and suppresses horizontally aligned regions exhibiting the highest activation responses. By masking these dominant regions, the network is guided to attend to less salient but potentially discriminative areas, thereby alleviating the loss of critical information caused by occlusion and enhancing the robustness of the feature representation.

The low-response feature region retained after the Top DropBlock operation is further processed via Global Max Pooling (GMP) to generate G_{drop}, an implicit global feature with dimensions 512×1×1. Serving as a complementary representation to the explicit branch, this feature enhances the completeness and robustness of the overall feature embedding, thereby facilitating global modeling with improved generalization capability.

3.4 Dual-path Local Branch

As illustrated in Fig. 6, the dual-path local branch consists of two integral components: the local feature reorganization path and the local feature concatenation path. These complementary pathways collaboratively enhance local feature representations by addressing fine-grained structural modeling and category discriminability. This architectural design effectively mitigates performance degradation resulting from missing or corrupted local information in occluded scenarios. Specifically, the local feature reorganization path introduces controllable information redundancy and enforces spatial continuity constraints to reinforce fine-grained feature learning. Concurrently, the local feature concatenation path emphasizes the enhancement of inter-region category separability, thereby augmenting the discriminative capability of local features. Functionally complementary and synergistic, these sub-paths jointly construct robust and semantically enriched local

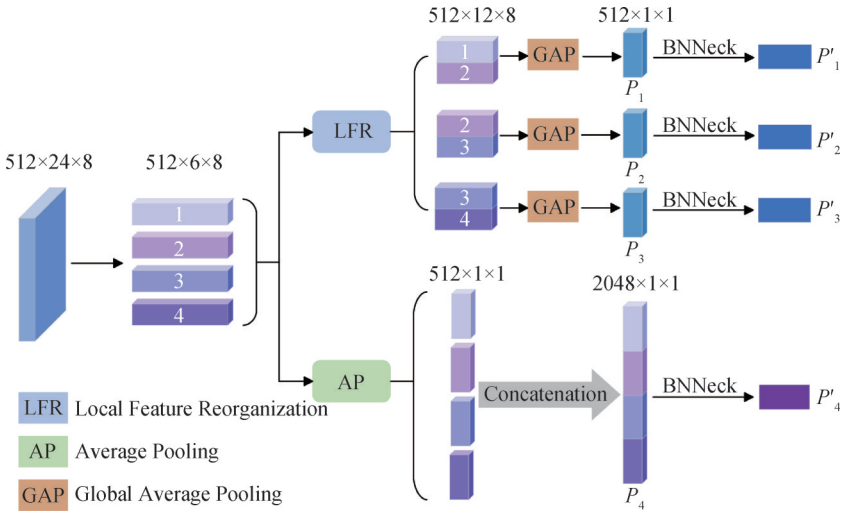

Fig.6 Network structure of the dual-path local branch

representations, significantly improving the model's regional perception and recognition accuracy under occlusion conditions.

In the subsequent sections, the local feature reorganization path and the local feature concatenation path will be introduced in detail, respectively.

(1) Local feature reorganization path

To alleviate the detrimental effects of local occlusions on Re-ID performance, a local feature reorganization strategy based on overlapping regions is proposed. In contrast to conventional methods that partition the feature map into uniformly spaced, non-overlapping segments, the proposed approach introduces spatial overlap between adjacent local regions. By embedding redundant information and enforcing continuity constraints across neighboring regions, this strategy enhances the model's sensitivity to fine-grained variations, thereby improving its robustness and efficacy under complex occlusion scenarios.

Specifically, the input feature map of dimension 512×24×8 is initially partitioned into four equal local regions along the horizontal axis. Subsequently, a sliding fusion strategy is employed to concatenate every two adjacent regions, producing three local regions with a 50% spatial overlap. Each overlapped region has dimensions of 512×12×8. This reorganization enables certain spatial locations to be shared between neighboring regions, thereby achieving redundant representations of discriminative information. The overlap design enhances the robustness against feature occlusion by ensuring that, even if one region is occluded, its critical features are partially preserved in adjacent regions. This mechanism effectively mitigates information loss due to occlusion, thereby improving the stability and robustness of the overall recognition process.

Finally, Global Average Pooling (GAP) is applied to the three reorganized local regions to achieve feature dimensionality reduction and global contextual aggregation. Following GAP, three fine-grained feature vectors P_1, P_2 and P_3, each of dimension 512×1×1, are obtained. These fine-grained vectors provide enhanced discriminative representations of local details, thereby contributing to improved accuracy and robustness during subsequent feature fusion.

(2) Local feature concatenation path

To address the degradation of discriminative capacity in local regions caused by occlusion, a local feature concatenation strategy is proposed. This approach compensates for information loss by aggregating semantic features from multiple local regions, thereby enhancing the overall discriminative power of the feature representation.

Firstly, the Average Pooling (AP) operation is applied to each of the four local regions obtained from the initial segmentation to obtain four local feature vectors, all with dimensions of 512×1×1. Subsequently, these four feature vectors are spliced along the channel dimensions to obtain a high-dimensional discriminative feature representation, P_4, which has dimensions of 2048×1×1. Through this concatenation strategy, the model is able to integrate information from different localized regions, thus enhancing the discriminative nature of the overall features and improving the recognition of occluded pedestrians.

3.5 Multi-objective Joint Loss Function

In this paper, we propose a novel multi-objective optimization framework that integrates complementary loss functions. As detailed in Section 2.3, the framework combines label-smoothed ID Loss, MS Loss, and Triplet Loss—each contributing distinct advantages—to jointly enhance the model's feature discriminability, robustness, and generalization capability. The overall formulation of the combined loss function is expressed as follows:

$$L_{total}=\lambda_1 L_{ID}+\lambda_2 L_{MS}+(1-\lambda_2) L_{Triplet} \tag{12}$$

Here, L_{ID}denotes the label-smoothed ID Loss, which is weighted by a balancing factorλ_1. The terms $L_{MS}$and $L_{Triplet}$represent the MS Loss and Triplet Loss,

respectively, both of which are constrained by a balancing factorλ_2.

Within the dual-path local branch, local features *(P_1–P_4)* are optimized exclusively using label-smoothed ID loss, intentionally excluding metric learning objectives. This design choice stems from the susceptibility of local regions to background occlusions and the presence of increased noise, which, if directly subjected to metric learning, may induce feature space contamination, hinder the acquisition of discriminative features, and potentially degrade overall model performance. Consequently, to ensure the stability and reliability of local feature representations, optimization is conducted solely under classification supervision.

During model training, the feature representation in unit space and rank space can be formally defined as two feature sets:

$$\begin{aligned} I &= \left\{ \mathrm{G}'_{\mathrm{drop}},\ \mathrm{P}'_1,\ \mathrm{P}'_2,\ \mathrm{P}'_3,\ \mathrm{P}'_4 \right\} \\ R &= \left\{ \mathrm{G}, \mathrm{G}_{\mathrm{drop}} \right\} \end{aligned} \tag{13}$$

where the set I denotes the feature space for supervised learning by label-smoothed ID loss only, and the set R denotes the feature space optimized using the joint loss function.

4 Experiments

To validate the effectiveness of the proposed method in occluded Re-ID tasks, comprehensive experiments were conducted on three benchmark datasets: Occluded-Duke, DukeMTMC-ReID, and Market-1501. The experimental procedure comprises the following steps:

(1)Firstly, the superiority of the multi-objective joint loss function over single-loss approaches is empirically evaluated in terms of feature discriminability and generalization capability.

(2)Secondly, ablation studies are performed on the visibility-guided positional attention mechanism, the global complementary branch, and the dual-path local branch to quantify the individual contributions of each component to overall performance enhancement. Furthermore, comparative experiments against state-of-the-art methods demonstrate that the proposed approach achieves competitive results across multiple key evaluation metrics.

(3)Finally, qualitative analysis is conducted to visualize the model's feature discrimination efficacy under occlusion conditions.

4.1 Datasets and Evaluation Measures

Comprehensive experiments are conducted on three widely adopted benchmark datasets: Occluded-Duke, DukeMTMC-ReID, and Market-1501. These datasets encompass various challenging conditions, enabling thorough evaluation of the model's generalization ability, cross-view matching robustness, and occlusion handling capability.

Occluded-Duke[12] is a benchmark dataset specifically constructed for occluded person Re-ID tasks, derived from the widely used DukeMTMC-ReID dataset and extended to emphasize occlusion scenarios observed in real-world surveillance environments. It comprises a total of 35,589 images corresponding to 1,812 pedestrian identities, and includes both holistic and partially occluded samples. The training set contains 702 identities (8, 010 images), while the test set includes 1, 110 identities, consisting of 2,210 query images and 17,761 gallery images. Representative examples from the Occluded-Duke dataset are presented in Fig. 7. The occlusion types predominantly include static obstacles and occlusions caused by non-target pedestrians.

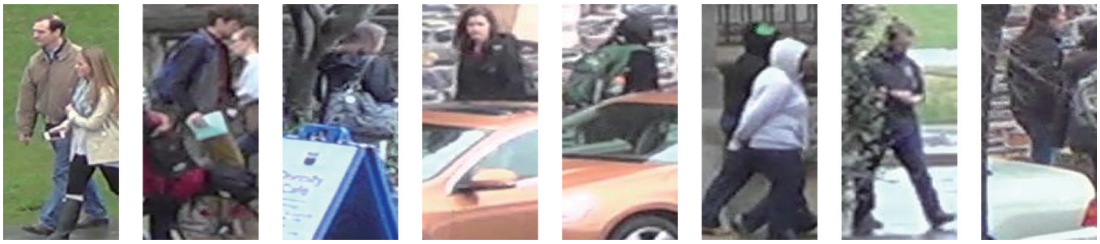

Fig.7 Example of the Occluded-Duke dataset

All query images and approximately 10% of the gallery images exhibit varying degrees of occlusion, ensuring that each matching set contains at least one occluded sample.

DukeMTMC-ReID[24] is captured on the Duke University campus using eight non-overlapping cameras covering a wide range of lighting, angle, background, and occlusion scenarios, and in particular includes a variety of occlusion scenarios to fully represent the difficulty of recognition in complex environments. The dataset contains 36, 411 images covering a total of 1, 812 identities. The training set consists of 702 pedestrians (16, 522 images), and the test set covers the remaining 1,110 pedestrians, containing 2, 228 query images and 17, 661 gallery images.

Market-1501[25] is one of the most widely used benchmark datasets in the field of Re-ID. The dataset provides a total of 32, 668 pedestrian images and is divided into non-overlapping training and test sets by identity, containing 751 pedestrians (12,936 images) and 750 pedestrians (19, 732 images), respectively. The test set is further divided into a query set (3,368 images) and

a gallery set for evaluating image retrieval performance.

To comprehensively evaluate the retrieval performance of the Re-ID model, two mainstream metrics are used: Cumulative Matching Characteristic (CMC) curve and mean Average Precision (mAP). The CMC measures the probability that the target pedestrian is successfully matched within the top-k retrieval results, with the Rank-1 accuracy being particularly critical. The mAP represents the mean of the Average Precision (AP) across all query samples, comprehensively reflecting both retrieval precision and recall. The formulas for CMC and mAP are as follows:

$$CMC(n)=\frac{1}{M}\sum_{i=1}^{M}p_i,\ n=1,2,\ldots,N \tag{14}$$

$$mAP=\frac{1}{C}\sum_{k=1}^{C}AP_k \tag{15}$$

here$M$is the sample number of pedestrian images, $N$is the total number of query images, if the ith image matches correctly, then p_i =1 otherwise p_i =0, C is the number of images to be recognized.

4.2 Implementation Details

In this paper, OSNet pre-trained on ImageNet is used as the baseline network. In the training stage, the input images are first channel-by-channel normalized (zero mean, unit variance) and uniformly adjusted to a spatial resolution of 384×128. In order to improve the generalization ability of the model, data enhancement strategies such as width-height scaling to 105% of the original size, random cropping, and horizontal flipping with 0.5 probability are used. Furthermore, Random Erasing[26] is introduced for occlusion augmentation, which randomly selects rectangular regions on images for occlusion to simulate diverse occlusion patterns, thereby improving the model's robustness in occluded scenarios.

The training process consists of 160 epochs with a batch size of 48, where each batch contains 8 pedestrian identities and 6 images per identity. The Adam optimizer is employed with momentum factors β_1 =0.9 and β_2 = 0.999, using an initial learning rate of 6×10^{-4} and a warm-up cosine annealing schedule[27] : During the first 10 epochs, the learning rate linearly increases from 6×10^{-5} to 6×10^{-4}, for the remaining 150 epochs, it gradually decays following a cosine function. The model is trained under the supervision of a multi-objective joint loss function to enhance feature discriminability and robustness. During testing, the final representation is obtained by concatenating dual-branch features, with cosine distance used for matching evaluation. Detailed experimental environment parameters are shown in Table 1.

4.3 Loss Function Validity Analysis

The multi-objective joint loss function combines the label-smoothed ID loss with two types of metric losses—MS Loss and Triplet Loss—to synergistically enhance both the discriminability and robustness of pedestrian features. To balance the contributions of different loss components, two hyperparameters, λ_1 and λ_2, are introduced: λ_1 controls the relative weights of the classification and metric losses (empirically set to 0.5), while λ_2 adjusts the influence between the MS Loss and the Triplet Loss, and is tuned experimentally.

Table 1 Experimental environment parameters

Configuration	Parameter
Operating System	ubuntu20.04
CPU	16 vCPU Intel(R) Xeon(R) Platinum 8481C
Memory	32GB
GPU	NVIDIA RTX4090D
Software Platform	Python3.8、PyTorch2.0.0、CUDA11.8

Table 2 presents an ablation study analyzing the impact of λ_2 on model performance using the Occluded-Duke dataset. When λ_2 = 0, only ID Loss and Triplet Loss are introduced; when λ_2 = 1, it is a combination of ID Loss and MS Loss. Fig. 8 illustrates a line graph of the effect of different λ_2 values on the model performance on the Occluded-Duke dataset. As can be seen from the figure, both Rank-1 and mAP metrics show an increasing trend as the value of λ_2 increases. In particular, when λ_2 = 0.9, the model reaches optimal performance, and the Rank-1 accuracy and mAP are improved to 64.3% and 57.1%, respectively. The experimental results show that the appropriate deployment of the weights of MS Loss and Triplet Loss significantly improves the discriminative ability of the model, verifying the positive effect of the multi-loss synergistic mechanism on the performance of the model, and further corroborating the complementary nature of different loss functions in terms of optimization objectives and learning preferences. Based on this, λ_2 = 0.9 is adopted as the optimal parameter for all experiments on the Occluded-Duke dataset in this paper.

Table 2 Experimental results for different λ_2 on Occluded-Duke dataset

Method		Occluded-Duke	
λ_2	$1-\lambda_2$	Rank-1	mAP
0	1.0	59.6	53.7
0.1	0.9	60.5	54.1
0.3	0.7	61.7	54.5
0.5	0.5	61.6	55.1
0.7	0.3	64.1	56.2
0.9	0.1	**64.3**	**57.1**
1.0	0	63.1	56.0

Table 3 provides a comprehensive comparative analysis of the sensitivity of the parameter λ_2 across DukeMTMC-ReID and Market-1501 datasets. As shown in Fig. 9, variations in λ_2 exert a moderate yet non-

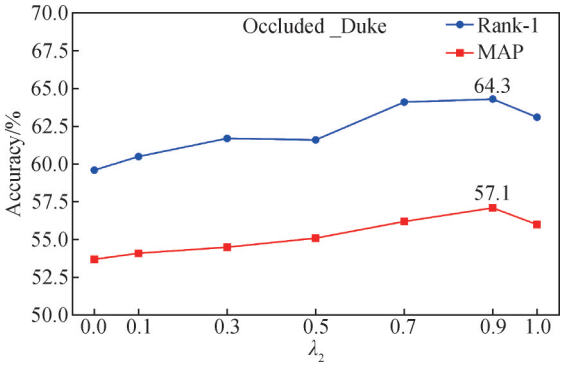

Fig.8 Impact of λ_2 variation on model performance on Occluded-Duke dataset

negligible impact on the overall model performance, influencing both Rank-1 accuracy and mAP. A detailed quantitative evaluation indicates that the model achieves its best performance when λ_2=0.7. Specifically, on the DukeMTMC-ReID dataset, the configuration yields 91.7% Rank-1 accuracy and 83.1% mAP, while on the Market-1501 dataset, it attains 96.4% Rank-1 accuracy and 91.0% mAP. These results demonstrate that an appropriately chosen λ_2 can effectively balance the contributions of different loss components, thereby enhancing both discriminative capability and generalization. Consequently, we adopt λ_2=0.7 as the optimal parameter setting for all subsequent experiments on these benchmark datasets.

Table 3 Experimental results for different λ_2 on DukeMTMC-ReID and Market-1501 datasets

Method		DukeMTMC-ReID		Market-1501	
λ_2	$1-\lambda_2$	Rank-1	mAP	Rank-1	mAP
0	1.0	90.9	82.5	96.4	90.9
0.1	0.9	90.7	82.4	96.2	90.9
0.3	0.7	91.3	82.9	96.4	90.8
0.5	0.5	91.1	82.6	96.4	90.9
0.7	0.3	**91.7**	**83.1**	**96.4**	**91.0**
0.9	0.1	91.4	82.8	96.2	90.9
1.0	0	91.0	82.6	96.2	90.6

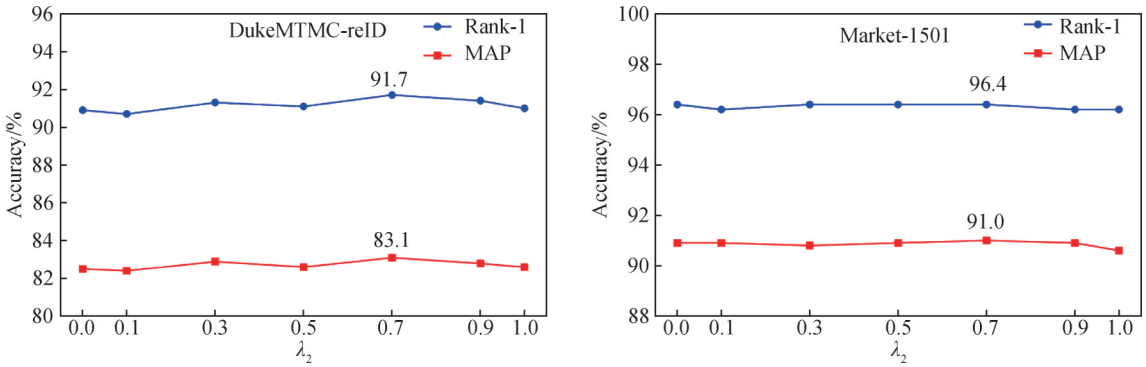

Fig.9 Impact of λ_2 variation on model performance on DukeMTMC-ReID and Market-1501 datasets

4.4 Ablation Study

To evaluate the effectiveness of the proposed VGDNet, we conduct systematic ablation studies across three benchmark datasets: Occluded-Duke, DukeMTMC-ReID, and Market-1501. As detailed in Tables 4 and 5, the experimental design incorporates six distinct model configurations: (1) Baseline: the base model; (2) Baseline+G: baseline with only the global complementary branch; (3) Baseline+P: baseline with only the dual-path local branch; (4) Baseline+G+P: baseline integrating both global and local branches; (5) Baseline+G+PAM: baseline combining global branch with positional attention mechanism; (6) Baseline+P+PAM: baseline combining local branch with positional attention mechanism; and (7)

Baseline+G+P+PAM: the complete VGDNet architecture incorporating all components.

The experimental results validate that VGDNet achieves substantial performance gains on Re-ID tasks, where each component module demonstrates both independent efficacy and synergistic benefits when integrated. The detailed analysis is as follows:

Table 4 Ablation study results of module combinations on Occluded-Duke dataset

Network Branch	Occluded-Duke	
	Rank-1	mAP
Baseline	57.4	44.9
Baseline+G	58.9	50.3
Baseline+P	60.4	52.4
Baseline+G+P	61.5	55.0
Baseline+G+PAM	62.2	55.7
Baseline+P+PAM	61.3	52.7
Baseline+G+P+PAM	**64.3**	**57.1**

The experimental results demonstrate consistent performance improvements across all three benchmark datasets through the incorporation of proposed modules,

Table 5 Ablation study results of module combinations on DukeMTMC-ReID and Market-1501 datasets

Network Branch	DukeMTMC-ReID		Market-1501	
	Rank-1	mAP	Rank-1	mAP
Baseline	88.6	73.5	94.8	84.9
Baseline+G	89.5	78.9	95.3	89.2
Baseline+P	90.4	80.0	95.9	89.5
Baseline+G+P	91.3	82.8	96.2	91.0
Baseline+G+PAM	89.8	79.4	95.4	89.2
Baseline+P+PAM	91.0	81.0	95.6	89.5
Baseline+G+P+PAM	**91.7**	**83.1**	**96.4**	**91.0**

with both the global complementary branch (G) and dual-path local branch (P) yielding significant enhancements over the baseline model. Particularly noteworthy is the superior performance of the dual-path local branch on the heavily occluded Occluded-Duke dataset, where it achieves remarkable improvements of 3% in Rank-1 accuracy and 7.5% in mAP compared to the baseline, conclusively validating its exceptional capability in addressing occlusion challenges while maintaining competitive performance on standard datasets.

When both global complementary branch and dual-path local branch (G+P) are integrated, the model achieves optimal performance, validating the complementary characteristics of the two in terms of feature representation. After further introducing PAM, the model achieves optimal performance on all three datasets. In the Occluded-Duke dataset, Rank-1 and mAP reach 64.3% and 57.1%, respectively, representing improvements of 6.9% and 12.2% over the baseline. on the DukeMTMC-reID dataset, Rank-1 and mAP reached 91.7% and 83.1%, respectively, representing improvements of 3.1% and 9.6% over the Baseline; on the Market-1501 dataset, Rank-1 and mAP reached 96.4% and 91.0%, respectively, representing improvements of 1.6% and 6.1% over the Baseline. This demonstrates that PAM effectively enhances the model's ability to perceive occluded regions, thereby improving overall recognition robustness.

4.5 Comparison with State-of-the-Art Method

Table 6 presents a comprehensive performance comparison between the proposed method and existing state-of-the-art approaches for occlusion handling in person Re-ID tasks. The experimental results demonstrate significant improvements across all evaluation metrics. Specifically, our method achieves a 12.9% and 19.8% enhancement in Rank-1 accuracy and mAP respectively compared to PGFA[12] , while outperforming HOreID[29] by 8.5% (Rank-1) and 13.3% (mAP). Even when compared with the competitive PGFL-KD[32] approach, our method maintains superior performance with gains of 1.3% and 3.0% in Rank-1 and mAP metrics. Furthermore, compared with recent representative Transformer-based method TransReID, the Rank-1 and mAP are improved by 0.1% and 1.4%, respectively. These consistent improvements across multiple benchmarks substantiate that VGDNet delivers enhanced recognition capability and superior robustness in complex occlusion scenarios.

Table 6 Comparison results with State-of-the-Art methods on Occluded-Duke dataset

Method	Occluded-Duke	
	Rank-1	mAP
Part-Aligned[28]	28.8	20.2
PCB[5]	42.6	33.7
PGFA[12]	51.4	37.3
PVPM[13]	47.0	37.7
HOreID[29]	55.8	43.8
Prit[30]	60.0	50.9
OAMN[31]	62.6	46.1
PGFL-KD[32]	63.0	54.1
PAT[33]	**64.5**	53.6
TransReID[34]	64.2	55.7
Ours	**64.3**	**57.1**
Ours(RK)[35]	**68.0**	**67.2**

Furthermore, to comprehensively evaluate the performance of the proposed network, this paper categorizes the methods into four categories: global feature methods, local feature methods, attention mechanism methods, and other mainstream methods. Comparative experiments are conducted on two mainstream Re-ID benchmark datasets: DukeMTMC-ReID and Market-1501. Table 7 presents the performance comparison results of the proposed method with various state-of-the-art methods on the aforementioned datasets. On the DukeMTMC-ReID dataset, the Rank-1 and mAP are 91.7% and 83.1%, respectively; on the Market-1501 dataset, the Rank-1 and mAP are 96.4% and 91.0%, respectively, outperforming most current state-of-the-art methods. Compared with the similarly multi-branch structure PLR-OSNet[46] , our method improves the discriminative power of local features, achieving a 0.1% improvement in Rank-1 and a 1.9% improvement in mAP on the DukeMTMC-ReID dataset, and a 0.8% improvement in Rank-1 and a 2.1% improvement in mAP on the Market-1501 dataset. These results further validate the effectiveness and robustness of the proposed method in cross-scene Re-ID tasks.

In addition, compared with recent high-performance methods such as PLIP and TransReID, although these approaches achieve higher accuracy by introducing large-

Table 7 Comparison results with State-of-the-Art methods on DukeMTMC-ReID and Market1501 datasets

Method Type	Method	DukeMTMC-reID		Market-1501	
		Rank-1	mAP	Rank-1	mAP
Global Method	SVDNet[36]	76.7	56.8	82.3	62.1
	BoT[19]	86.4	76.4	94.5	85.9
	Self-supervised[37]	89.0	78.2	94.7	86.7
	OSNet[22]	88.6	73.5	94.8	84.9
Local Method	SCPNet[38]	80.3	62.9	91.2	75.2
	PCB+RPP[5]	82.9	68.5	93.1	81.0
	HPM[6]	86.6	74.3	94.2	82.7
	ISP[39]	89.6	80.0	95.3	88.6
	MGN[40]	88.7	78.4	95.7	86.9
Attention Mechanism	Mancs[41]	84.9	71.8	93.1	82.3
	HA-CNN[17]	80.5	63.8	91.2	75.7
	AANet[42]	87.7	74.3	93.9	83.4
	CASN[43]	87.7	73.7	94.4	82.8
	ABDNet[44]	89.0	78.6	95.6	88.3
Other Method	BDB[45]	86.8	72.1	94.2	84.3
	PLR-OSNet[46]	91.6	81.2	95.6	88.9
	GPS[47]	88.2	78.7	95.2	87.8
	DeepMiner[48]	91.2	81.8	95.7	90.4
	PLIP[49]	91.1	81.7	**96.8**	**91.4**
	TransReID[34]	90.7	82.0	95.2	88.9
	Ours	91.7	83.1	96.4	91.0
	Ours(RK)[35]	**93.9**	**90.3**	**96.9**	**94.9**

scale pre-trained models or Transformer-based architectures (e. g., PLIP achieves a Rank-1 accuracy of 96.8% and an mAP of 91.4% on Market-1501, while TransReID attains a Rank-1 accuracy of 95.2% and an mAP of 88.9%), they typically rely on substantial parameter sizes and computational costs, making them less efficient for practical deployment. In contrast, our proposed model maintains a much lower complexity while achieving comparable or even superior performance under occluded scenarios, which highlights its advantages in lightweight design and practical applicability.

4.6 Visual Analysis

To further validate the robustness of the proposed VGDNet under occlusion conditions, we conducted a comprehensive visualization analysis of its recognition performance. As illustrated in Fig.10-12, the visualization results demonstrate the model's retrieval capability, where black bounding boxes denote query images and "Top-10" represents the highest-ranked candidate matches based on similarity metrics. Correct identifications are highlighted with green bounding boxes, while erroneous matches are marked in red. This visual assessment provides empirical evidence of the model's discriminative power in occlusion scenarios.

Fig. 10 presents a comparative visualization of retrieval performance under occlusion scenarios in the Occluded-Duke dataset. The baseline network, relying solely on single-scale feature extraction, demonstrates notable performance degradation when handling occluded samples. In the first test case involving inter-person occlusion, the baseline fails to discriminate fine-grained features between the target pedestrian and occluding individuals, resulting in ambiguous feature representations. Consequently, its Top-10 retrieval results contain minimal correct matches. The second case examines object occlusion, where the baseline's inability to model local occlusion patterns leads to ineffective capture of discriminative semantic attributes (e. g., clothing texture), significantly compromising recognition accuracy. By contrast, our proposed framework—incorporating visibility-aware feature modulation and a dual-branch architecture—dynamically adjusts feature contributions from both global and local branches based on visibility estimation. This adaptive mechanism enhances feature representation in occluded regions while effectively suppressing occlusion-induced interference, as evidenced by the improved retrieval accuracy.

Fig.10 Visualization results of Person Re-Identification on Occluded-Duke dataset

Fig. 11 presents the person Re-ID performance on the DukeMTMC-ReID dataset. The results demonstrate consistently high recognition accuracy across various occlusion conditions, including inter-pedestrian and object occlusions. Fig. 12 further evaluates the proposed method on the Market-1501 dataset, confirming its robustness even in scenarios with minimal or no occlusion. These visualizations validate the generalizability of our approach under diverse occlusion patterns.

Fig.11 Visualization results of Person Re-Identification on DukeMTMC-ReID dataset

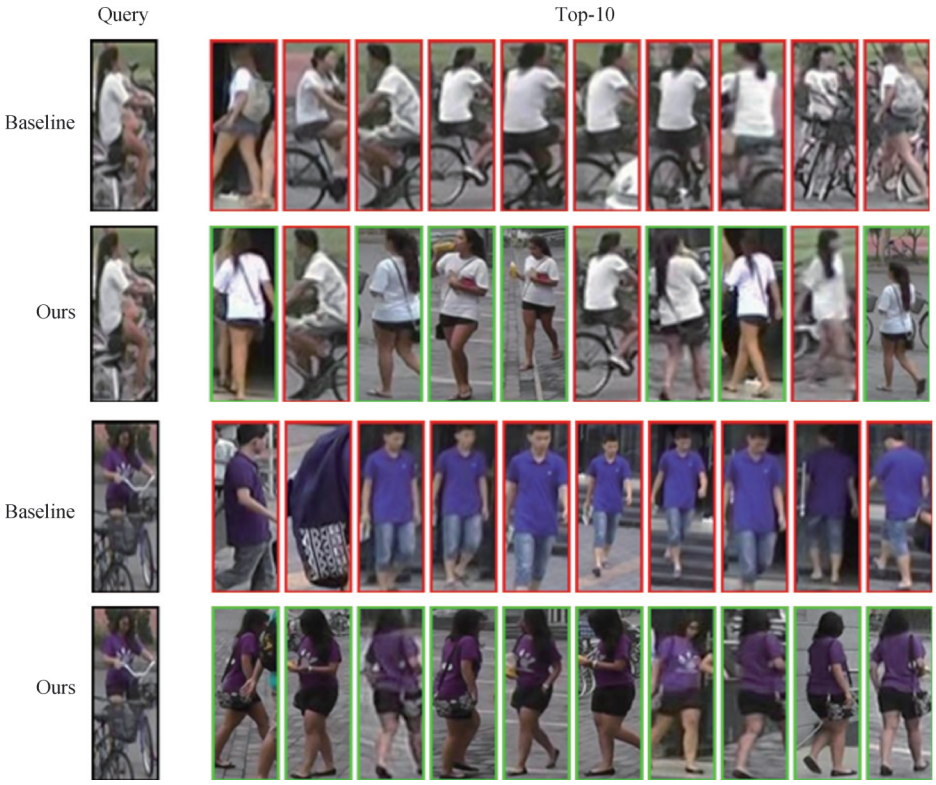

Fig.12 Visualization results of Person Re-Identification on Market-1501 dataset

Collectively, the visualization analysis substantiates that the proposed method consistently achieves robust recognition performance across varying occlusion complexities. These empirical results not only validate the method's efficacy but also demonstrate its practical viability for real-world deployment scenarios.

4.7 Lightweight Analysis

To evaluate the advantages of our method in lightweight design and its feasibility for practical deployment, we compare its parameter count and floating-point operations (FLOPs) against several mainstream convolutional neural networks. The detailed quantitative comparison results are presented in Table 8.

Table 8 Comparison of model parameters between mainstream networks and the proposed method.

network	Params(M)	FLOPs($\times 10^8$)
ResNet-50	23.5	26.7
ResNet-101	42.5	50.9
ResNet-152	58.1	75.6
OSNet	2.2	9.8
Ours	8.4	21.2

The proposed method does not rely on external auxiliary modules such as human pose estimation or generative adversarial networks, which effectively reduces the structural complexity and training difficulty of the model. While maintaining high recognition performance, the model achieves a good balance between lightweight design and computational efficiency. As shown in the table, the proposed network exhibits significantly lower complexity than mainstream deep backbones, with 8.4M parameters and 21.2×10^8 FLOPs. Compared with ResNet-50, the number of parameters is reduced by approximately 64%, and the computational cost decreases by about 21%. Although the parameter scale of our model is slightly larger than that of OSNet, it remains much more lightweight than the deeper ResNet-101 and ResNet-152, while achieving superior recognition performance under occluded scenarios. In addition, when tested on an NVIDIA RTX 4090D GPU with a batchsize of 1 and input resolution of 384×128 pixels, our model attains an average inference speed of about 65 FPS, demonstrating high inference efficiency and strong potential for practical deployment in real-time person re-identification applications such as intelligent surveillance and mobile devices.

5 Conclusion

In this paper, we propose a robust recognition network that integrates visibility-aware cues into a dual-branch a- rchitecture to address the challenges of feature degradation caused by poor visibility in the occluded person Re-ID task. The proposed framework incorporates a positional attention mechanism to capture spatially informative regions within the image, while constructing a visibility factor derived from attention information entropy to dynamically regulate the balance between

global and local feature representation. Specifically, the global complementary branch augments semantic representation through synergistic explicit-implicit feature integration coupled with a Top DropBlock strategy. Concurrently, the dual-path local branch enhances fine-grained structural representation via region reorganization and high-dimensional feature concatenation operations. Furthermore, a multi-objective joint loss function is formulated to facilitate synergistic optimization of both discriminative capability and generalization performance. Comprehensive experiments conducted on standard benchmarks including Occluded-Duke, DukeMTMC-ReID, and Market-1501 demonstrate that the proposed approach achieves state-of-the-art performance and robust generalization under occlusion scenarios. Nevertheless, this method still has certain limitations. Under extreme occlusion, when a large portion of the identity-relevant regions is covered, the lack of effective identity cues may still lead to performance degradation. In addition, when the input image is of low resolution, the loss of local region details can weaken the fine-grained feature modeling ability of the local branch. In future work, we will explore enhancing feature completion mechanisms and improving the model's adaptability to low-resolution inputs to further strengthen its robustness in more complex real-world scenarios.

Author Contribution:

Menghan An and Yanyan Zhang conceived and designed the study; Yanyan Zhang supervised the study; Menghan An performed the methodology and experiments, and wrote the original draft; Yu Qin performed the data analysis and visualization; Yanyan Zhang reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding Information:

This research received no external funding.

Data Availability:

The authors declare that the main data supporting the findings of this study are available within the paper and its Supplementary Information files.

Conflicts of Interest:

The authors declare no competing interests.

Dates:

Received 11 August 2025; Accepted 16 March; Published online 31 March 2026

References

[1] Chen W, Xu X, Jia J, et al. (**2023**). Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2023**: 15050-15061.

[2] Fu D, Chen D, Yang H, et al. (**2022**). Large-scale pre-training for person re-identification with noisy labels[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2022**: 2476-2486.

[3] Luo H, Wang P, Xu Y, et al. (**2021**). Self-supervised pre-training for transformer-based person re-identification[J]. *arXiv preprint arXiv:* 2111.12084.

[4] Li D, Chen S, Zhong Y, et al. (**2022**). Dip: Learning discriminative implicit parts for person re-identification[J]. *arXiv preprint arXiv:* 2212.13906.

[5] Sun Y, Zheng L, Yang Y, et al. (**2018**). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline) [C]. in Proc. European Conf. on Computer Vision (ECCV), **2018**: 480-496.

[6] Fu Y, Wei Y, Zhou Y, et al. (**2019**). Horizontal pyramid matching for person re-identification[C]. in Proc. AAAI Conf. on Artificial Intelligence (AAAI), **2019**, 33(01): 8295-8302.

[7] He L, Liang J, Li H, et al. (**2018**). Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach[C]. in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), **2018**: 7073-7082.

[8] Wang Z, Zhu F, Tang S, et al. (**2022**). Feature erasing and diffusion network for occluded person re-identification[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2022**: 4754-4763.

[9] Zhao C, Lv X, Dou S, et al. (**2021**). Incremental generative occlusion adversarial suppression network for person ReID [J]. *IEEE Transactions on Image Processing* , **2021**, 30: 4212-4224.

[10] Qian X, Fu Y, Xiang T, et al. (**2018**). Pose-normalized image generation for person re-identification[C]. in Proc. European Conf. on Computer Vision (ECCV), **2018**: 650-667.

[11] Zhang S, Ji M, Li Y, et al. (**2024**). Imagine the U-nseen: Occluded Pedestrian Detection via Adversa-rial Feature Completion[J]. *arXiv preprint arXiv:* 2405.01311.

[12] Miao J, Wu Y, Liu P, et al. (**2019**). Pose-guided feature alignment for occluded person re-identification[C]. in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), **2019**: 542-551.

[13] Gao S, Wang J, Lu H, et al. (**2020**). Pose-guided visible part matching for occluded person reid[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2020**: 11744-11752.

[14] Somers V, De Vleeschouwer C, Alahi A. (**2023**). Body part-based representation learning for occluded person re-identification[C]. in Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV), **2023**: 1613-1623.

[15] Fu J, Liu J, Tian H, et al. (**2019**). Dual attention network for scene segmentation[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2019**: 3146-3154.

[16] Zhang Z, Lan C, Zeng W, et al. (**2020**). Relation-aware global attention for person re-identification[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2020**: 3186-3195.

[17] Li W, Zhu X, Gong S. (**2018**). Harmonious attention network for person re-identification[C]. in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), **2018**:

2285-2294.

[18] Fan Y, Gong X, He Y. (**2023**). DSF-net: occluded person re-identification based on dual structure features[J]. *Neural Computing and Applications* , **2023**, 35(4): 3537-3550.

[19] Luo H, Gu Y, Liao X, et al. (**2019**). Bag of tricks and a strong baseline for deep person re-identification[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), **2019**: 1487-1495.

[20] Wang X, Han X, Huang W, et al. (**2019**). Multi-similarity loss with general pair weighting for deep metric learning[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2019**: 5022-5030.

[21] Hermans A, Beyer L, Leibe B. (**2017**). In defense of the triplet loss for person re-identification[J]. *arXiv preprint arXiv:* 1703.07737.

[22] Zhou K, Yang Y, Cavallaro A, et al. (**2019**). Omni-scale feature learning for person re-identification[C]. in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), **2019**: 3702-3712.

[23] Quispe R, Pedrini H. (**2021**). Top-db-net: Top dropblock for activation enhancement in person re-identification[C]. 2020 25th International conference on pattern recognition (ICPR), IEEE, **2021**: 2980-2987.

[24] Zheng Z, Zheng L, Yang Y. (**2017**). Unlabeled samples generated by gan improve the person re-identification baseline in vitro[C]. in Proc. IEEE Int. Conf. on Computer Vision (ICCV), **2017**: 3754-3762.

[25] Zheng L, Shen L, Tian L, et al. (**2015**). Scalable person re-identification: A benchmark[C]. in Proc. IEEE Int. Conf. on Computer Vision (ICCV), **2015**: 1116-1124.

[26] Zhong Z, Zheng L, Kang G, et al. (**2020**). Random erasing data augmentation[C]. in Proc. AAAI Conf. on Artificial Intelligence (AAAI), **2020**,34(07): 13001-13008.

[27] Loshchilov I, Hutter F. (**2016**). Sgdr: Stochastic gradient descent with warm restarts[J]. *arXiv preprint arXiv:* 1608.03983.

[28] Zhao L, Li X, Zhuang Y, et al. (**2017**). Deeply-learned part-aligned representations for person re-identification[C]. in Proc. IEEE Int. Conf. on Computer Vision (ICCV), **2017**: 3219-3228.

[29] Wang G, Yang S, Liu H, et al. (**2020**). High-order information matters: Learning relation and topology for occluded person re-identification[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2020**: 6449-6458.

[30] Ma Z, Zhao Y, Li J. (**2021**). Pose-guided inter-and intra-part relational transformer for occluded person re-identification [C]. Proceedings of the 29th ACM international conference on multimedia, **2021**: 1487-1496.

[31] Chen P, Liu W, Dai P, et al. (**2021**). Occlude them all: Occlusion-aware attention network for occluded person re-id [C]. in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), **2021**: 11833-11842.

[32] Zheng K, Lan C, Zeng W, et al. (**2021**). Pose-guided feature learning with knowledge distillation for occluded person re-identification[C]. Proceedings of the 29th ACM international conference on multimedia, **2021**: 4537-4545.

[33] Li Y, He J, Zhang T, et al. (**2021**). Diverse part discovery: Occluded person re-identification with part-aware transformer [C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2021**: 2898-2907.

[34] He S, Luo H, Wang P, et al. Transreid: Transformer-based object re-identification[C]. in Proc. IEEE Int. Conf. on Computer Vision (ICCV), **2021**: 15013-15022.

[35] Zhong Z, Zheng L, Cao D, et al. (**2017**). Re-ranki-ng person re-identification with k-reciprocal enco-ding[C]. in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), **2017**: 1318-1327.

[36] Sun Y, Zheng L, Deng W, et al. (**2017**). Svdnet for pedestrian retrieval[C]. in Proc. IEEE Int. Conf. on Computer Vision (ICCV), **2017**: 3800-3808.

[37] Chen F, Wang N, Tang J, et al. (**2021**). A feature disentangling approach for person re-identification via self-supervised data augmentation[J]. *Applied Soft Computing* , **2021**, 100: 106939.

[38] Zhao H, Tian M, Sun S, et al. (**2017**). Spindle net: Person re-identification with human body region guided feature decomposition and fusion[C]. in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), **2017**: 1077-1085.

[39] Zhu K, Guo H, Liu Z, et al. (**2020**). Identity-guided human semantic parsing for person re-identification[C]. European conference on computer vision, Cham: Springer International Publishing, **2020**: 346-363.

[40] Wang G, Yuan Y, Chen X, et al. (**2018**). Learning discriminative features with multiple granularities for person re-identification[C]. Proceedings of the 26th ACM international conference on Multimedia, **2018**: 274-282.

[41] Wang C, Zhang Q, Huang C, et al. (**2018**). Mancs: A multi-task attentional network with curriculum sampling for person re-identification[C]. in Proc. European Conf. on Computer Vision (ECCV), **2018**: 365-381.

[42] Tay C P, Roy S, Yap K H. (**2019**). Aanet: Attribute attention network for person re-identifications[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2019**: 7134-7143.

[43] Zheng M, Karanam S, Wu Z, et al. (**2019**). Re-ide-ntification with consistent attentive siamese netwo-rks[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2019**: 5735-5744.

[44] Chen T, Ding S, Xie J, et al. (**2019**). Abd-net: Att-entive but diverse person re-identification[C]. in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), **2019**: 8351-8361.

[45] Dai Z, Chen M, Gu X, et al. (**2019**). Batch dropbl-ock network for person re-identification and beyo-nd[C]. in Proc. IEEE/CVF Int. Conf. on Comput-er Vision (ICCV), **2019**: 3691-3701.

[46] Xie B, Wu X, Zhang S, et al. (**2020**). Learning diverse features with part-level resolution for person re-identification [C]. Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Cham: Springer International Publishing, **2020**: 16-28.

[47] Nguyen B X, Nguyen B D, Do T, et al. (**2021**). Graph-based person signature for person re-identifi-cations[C]. in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), **2021**: 3492-3501.

[48] Benzine A, Seddik M E A, Desmarais J. (**2021**). Deep miner: a deep and multi-branch network whi-ch mines rich and diverse features for person re -identification[J]. *arXiv preprint arXiv:* 2102.09321.

[49] Zuo J, Hong J, Zhang F, et al. (**2024**). Plip: Language-image pre-training for person representation learning[J]. *Advances in Neural Information Processing Systems, 37* , 45666-45702.