

Article

Scene-Prompt-Driven Dynamic Routing Expert Network for Open-World Person Re-Identification

Peng Dong , Hongbin Liu * , Xiuyi Guo , Jilong Li , Yitong Zhou  and Baoxu Wang 

School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan 250101, China; 2024080105@stu.sdjzu.edu.cn (P.D.); 2024080120@stu.sdjzu.edu.cn (X.G.); 2025080110@stu.sdjzu.edu.cn (J.L.); 2025085137@stu.sdjzu.edu.cn (Y.Z.); 2025085123@stu.sdjzu.edu.cn (B.W.)

* Correspondence: liuhongbin19@sdjzu.edu.cn

Abstract

Open-world person re-identification (ReID) faces severe spatially asymmetric interferences, such as partial occlusion and illumination distortion. Existing models adopting static weight fusion irreversibly corrupt identity representations when processing localized noise. To address this, we propose a Scene-Prompt-Driven Dynamic Routing Expert Network (SPDR-Net). Guided by spatial topological prior constraints, SPDR-Net treats person semantic parts as independent local experts. First, a semantic expert module decouples and refines local features using human spatial topology priors. Second, a Prompt-Guided Dynamic Routing (PGDR) network extracts high-level scene contexts to dynamically evaluate each expert's reliability, assigning routing weights to attenuate noise propagation. Finally, a global-local fusion module superimposes high-purity local features onto a global identity anchor, maintaining topological integrity to generate a unified descriptor for robust person re-identification. Extensive experiments on seven public benchmarks and a newly constructed real-world street dataset (SD-ReID) demonstrate that SPDR-Net achieves highly competitive performance against state-of-the-art methods and exhibits superior robustness under severe spatial-asymmetric interferences. Furthermore, end-to-end multi-camera closed-loop tests verify its robust decision-making capability and high engineering application value in real-world security systems.

Keywords: person re-identification; dynamic routing; mixture of experts; spatially asymmetric interference; scene prompt learning

MSC: 68U10



Academic Editor: Jonathan Blackledge

Received: 8 June 2026

Revised: 19 July 2026

Accepted: 20 July 2026

Published: 31 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Person re-identification (ReID) serves as a core component of visual decision support systems in smart cities [1]. This technology accurately retrieves and matches specific target pedestrians across monitoring camera networks with non-overlapping fields of view. With the deep integration of smart cities and public security monitoring, the research focusing on ReID technology has shifted from controlled scenes to open scenes with high uncertainty. However, ReID technology in the open world faces more severe and complex challenges than laboratory datasets: (1) Open scenes possess environmental randomness. Target images are often collected under composite interference, such as uncontrolled illumination changes, weather conditions, and low-resolution imaging caused by long distances [2]. This unpredictable domain shift phenomenon degrades the feature extraction

ability of models trained on high-quality static images. (2) Closed-world ReID usually assumes that the query target exists in the gallery and the identity categories are fixed. However, in real open-world retrieval, the monitoring gallery is often mixed with a large number of unknown and similar non-target persons. Meanwhile, over time, the monitoring system needs to continuously address new identities and appearance changes that have never appeared in the training stage [3,4]. This requirement imposes stringent demands on the cross-domain generalization and high-level semantic discrimination ability of the ReID model. (3) Existing studies usually regard occlusion, cross-modality, or resolution degradation as global domain shifts. However, in the open world, this environmental noise often manifests as spatial asymmetry. Specifically, environmental interference on persons is not uniformly distributed in space. For example, roadside non-motorized vehicles may only occlude the lower half of the person, causing an asymmetric loss of spatial features [5,6]. Strong local top lighting or shadows may only cause modal changes in the upper half of the body. As shown in Figure 1, this spatially asymmetric interference heavily degrades traditional static fusion models. Because these conventional models treat all body parts equally, their identity representations are easily dominated by local noise features.

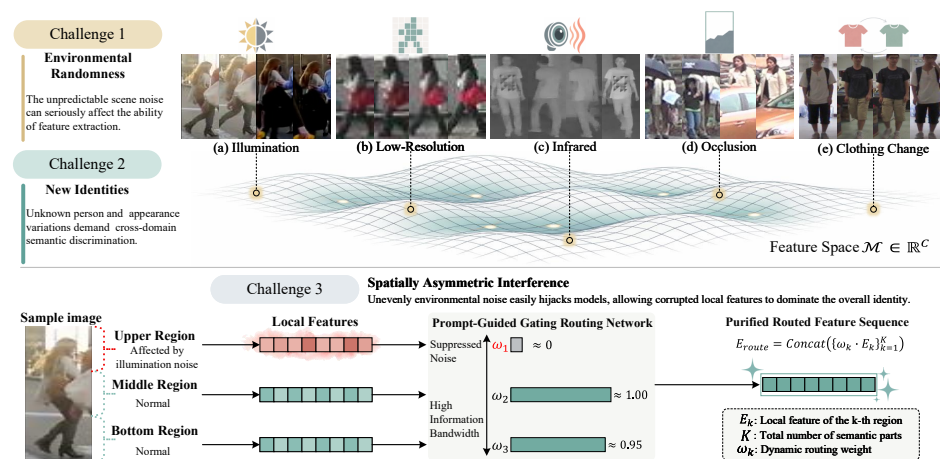


Figure 1. Challenges in open-world person re-identification. The upper part shows the composite challenges brought by environmental randomness (such as sudden illumination changes, low resolution, and infrared cross-modality) and target appearance mutations (such as severe occlusion and frequent clothing changes). The lower part shows the phenomenon of spatially asymmetric interference and its countermeasure. For a sample image polluted by local noise, traditional static feature fusion models are easily dominated by this local noise. This dominance leads to misidentification. In contrast, our SPDR-Net can accurately perceive the spatial asymmetry of environmental interference. Through non-linear reasoning, this network reduces the feature weights of the damaged parts. Consequently, it mitigates the propagation of local noise and finally outputs a high-purity routing feature sequence to achieve robust identity recognition. The arrows indicate the feature-processing direction; the colored token blocks denote features from different body regions; gray or attenuated blocks indicate suppressed noisy features; and E_k , K , and ω_k denote the k -th local feature, the number of semantic parts, and the corresponding dynamic routing weight, respectively.

To address these challenges, researchers have proposed many solutions. One category is Vision Transformer-based methods, represented by TransReID. These methods attempt to introduce camera IDs or posture information as auxiliary biases [7]. Another category is prompt learning-based methods, such as VersReID [8] or CLIP-ReID [9]. These prompt-based methods utilize the semantic prior of large language models to assist feature alignment [10]. Although these methods have achieved breakthroughs on specific datasets, they still have two limitations when handling the aforementioned “spatially asymmetric interference”. (1) Redundancy of global priors. Existing prompt learning methods mostly

apply scene information to global feature alignment. Therefore, these methods cannot finely evaluate the reliability of different body parts under specific interference. (2) Static feature fusion. Part-aware models, represented by PCB [11] or MGN [12], usually adopt a fixed weight fusion strategy. When a specific local feature suffers from severe noise interference, this noisy feature will inevitably pollute the final identity representation. This pollution leads to poor robustness of the model in extreme scenes.

This paper argues that the core of solving the open-world ReID problem does not lie in arbitrarily increasing feature dimensions. Instead, the core lies in establishing a “dynamic trust mechanism” between environmental semantics and local features. We find that the human visual system naturally leverages spatial topology to adaptively shift attention. When the environment indicates that the lower body is occluded, the attention bandwidth is dynamically redirected to the visible head and torso regions. This decision process based on environment prompts is exactly an area rarely covered by existing models.

Inspired by this cognitive mechanism, the paper proposes a Scene-Prompt-Driven Dynamic Routing Expert Network (SPDR-Net) to complete ReID task. This method incorporates occluded or severely affected local features into a dynamically optimized Mixture of Experts (MoE) system. Specifically, we design a lightweight prompt-guided gating module. This gating module takes the scene prompt vector as a contextual reference [13]. It then dynamically assigns routing weights to spatially semantic local expert features. In this way, the model can autonomously suppress features from severely noise-polluted parts based on real-time environmental feedback. Meanwhile, the model enhances the features of clear and reliable parts.

The contributions of this paper are summarized as follows:

- (1) This paper proposes a Scene-Prompt-Driven Dynamic Routing Expert Network (SPDR-Net). The architecture extends the conventional global prompt alignment to fine-grained local expert routing. Furthermore, this architecture favorably enhances the scene adaptive capability of the model in open-world environments.
- (2) A sparsity routing entropy loss function is designed. By explicitly penalizing the information entropy of the routing weight distribution, this loss function forces the network to generate polarized weights under interference, thereby effectively suppressing the representations of noisy parts.
- (3) We construct a real-world street scene dataset named SD-ReID, which contains 147 identities and over 10,000 images. Moreover, we build a multi-scene real-world evaluation platform integrating software and hardware. This platform fills the gap of existing research in industrial closed-loop testing.
- (4) Extensive experiments are conducted on seven mainstream public datasets and the newly constructed SD-ReID dataset (covering normal, occlusion, clothing-changing, and infrared scenes). The experimental results show that the SPDR-Net significantly outperforms current methods in cross-scene generalization ability.

2. Related Work

2.1. Open-World Person Re-Identification

Early research on person re-identification (ReID) mainly focused on closed-world assumptions in controlled environments. However, Zheng et al. [14] pioneered the identification of the “Open-World” challenge in real-world surveillance networks. In these unconstrained scenarios, the gallery contains a massive number of unknown non-target distractors, and the target appearance evolves drastically over time. With the evolution of deep learning, Vision Transformer (ViT)-based architectures (such as TransReID [15]) have achieved breakthroughs in cross-domain feature extraction. This success relies on their powerful global receptive fields. Based on this foundation, researchers have pro-

posed numerous customized solutions to cope with the complex physical environments and target changes in the open world. For example, Chen et al. [16] recovered the lost details of distant targets through a cross-resolution joint learning strategy. Gao et al. [17] designed an identity-guided collaborative learning framework to address the frequent clothing-changing problem in cross-period tracking. In recent years, with the application of large foundation models, prompt learning has been widely introduced into the vision domain to solve multi-scene generalization problems. CLIP-ReID, proposed by Li et al. [9], pioneered the introduction of the vision–language model prompt mechanism into ReID tasks. This method significantly improves the representation ability of cross-domain features without additional annotations. Zheng et al. [8] further proposed the VersReID framework. This framework utilizes learnable pure-vision scene prompt vectors to dynamically adapt to various environmental domain shifts. More recently, Wen et al. proposed Prompt-Anchored Vision–Text Distillation (PAD) for lifelong person ReID, which uses prompt anchors to transfer vision–language knowledge while mitigating catastrophic forgetting across sequentially arriving ReID domains [18]. However, the above prompt learning-based methods usually treat environmental noise (such as illumination distortion and occlusion) as a kind of “global domain shift” [19]. Therefore, they attempt to align data distributions at the global feature level [20]. In the real open world, this strategy ignores the fact that environmental interference often possesses strong spatial asymmetry [21]. When severe corruption occurs in local image regions, relying solely on global prompts easily causes features in undamaged areas to be undesirably affected [22,23].

2.2. Mixture of Experts

To combat spatially asymmetric environmental noise (including physical occlusion, local illumination mutation, and local blur), researchers have turned their attention to human part-based local feature extraction networks. The PCB model proposed by Sun et al. pioneered local feature learning through a horizontal slicing strategy. Subsequently, Zeng et al. [24] explored the problem of local illumination adaptation to address visual loss caused by asymmetric interference. Wang et al. [25] proposed a Feature Erasing and Diffusion (FED) network to infer the semantics of occluded parts. Zhang et al. [5] and Mishra et al. [26] successively utilized the attention mechanism of Transformers for part perception and local denoising reconstruction. These local networks have alleviated the spatial feature alignment problem (Spatial Misalignment [27]) caused by various local feature asymmetries to some extent. Nevertheless, the vast majority of these methods still adopt the fusion strategy of static feature concatenation or average pooling. This static strategy means that whether a local feature is severely occluded or in extreme shadow, its polluted noise may still be retained in the final identity metric. To overcome this static bottleneck, the concept of the Mixture of Experts (MoE) system has begun to emerge in the vision domain. The V-MoE proposed by Riquelme et al. [28] proved that the sparse routing mechanism can effectively improve the anti-interference ability of large vision models. Puigcerver et al. [29] further explored the application of soft routing in image patch processing. Inspired by this, some recent studies attempt to introduce dynamic routing into fine-grained image retrieval and video ReID. These studies select different channels through a gating network [30,31].

Although the aforementioned local networks and routing mechanisms have achieved promising progress in their respective studies, most existing methods separate the inherent connection between macro environmental perception and micro local routing [32]. When facing local asymmetric disturbances in the open world, traditional dynamic routing often lacks high-level decision guidance with cross-domain prior knowledge [33]. To fill this research gap, the SPDR-Net framework proposed in this paper innovatively and deeply

integrates scene prompt learning with local Mixture of Experts (MoE) routing [34]. This framework no longer relies on blind static feature fusion; instead, it treats the extracted scene prompts as the macro decision context for dynamic routing [35]. Through the sparse gating mechanism, the network can autonomously perceive and strongly suppress expert paths with severe local contamination. This design provides a novel dynamic adaptive paradigm for dealing with local asymmetric disturbances in the open world [36].

2.3. Structural Differences from Existing Paradigms

To clarify our contributions, we explicitly distinguish the proposed SPDR-Net from existing part-based, Mixture-of-Experts (MoE), and prompt learning ReID methods. (1) Unlike traditional part-based models (e.g., PCB [11]) that rely on static, rigid feature concatenation under a fixed fusion weight, SPDR-Net introduces a dynamic gating routing network to adaptively adjust part weights based on real-time environmental context. (2) Unlike conventional vision MoE networks (e.g., V-MoE [28]) where sparse experts share the same semantic output space and are merged via a weighted sum, our spatial experts represent physically distinct, non-overlapping body parts. We adopt a spatial concatenation strategy to strictly preserve the human body's vertical topology. (3) Unlike typical prompt learning ReID frameworks (e.g., VersReID [8]) that utilize scene prompts to perform global feature alignment, SPDR-Net uses the extracted scene prompts as a macroscopic decision context to guide the fine-grained routing of local experts. This design bridges the gap between global environmental perception and local feature selection, providing a highly adaptive routing paradigm.

3. Proposed Method

3.1. Overview

The proposed SPDR-Net framework is a unified dynamic routing architecture designed to solve the person ReID problem under complex open-world environments. Its overall architecture is shown in Figure 2.

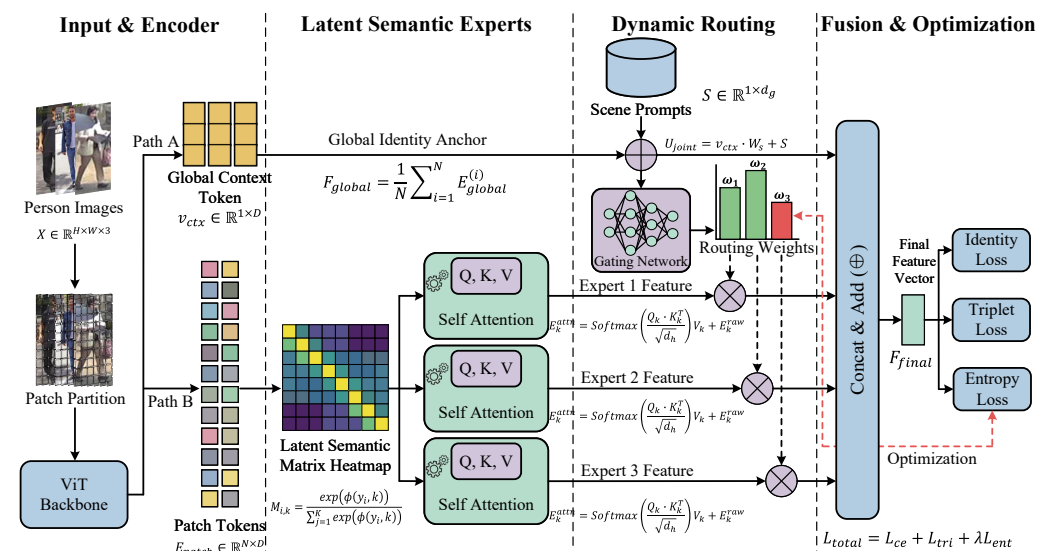


Figure 2. Overall architecture of the proposed SPDR-Net. **(Input and Encoder)** Input image X is fed into the ViT backbone to extract global spatial features, which are split into the global identity anchor F_{global} and the patch token sequence E_{patch} . **(Latent Semantic Experts)** Guided by human spatial topology priors, the patch sequence is decoupled into K latent semantic part experts $\{E_k\}_{k=1}^K$, which are independently denoised and refined via self-attention (Q, K, V) in their respective subspaces. **(Dynamic Routing)** The Prompt-Guided Dynamic Routing (PGDR) network extracts the global

visual context v_{ctx} , aligns it with the learnable scene prompt prior $S \in \mathbb{R}^{1 \times d_g}$ to form the decision basis U_{joint} , and infers polarized routing weights $\omega = [\omega_1, \dots, \omega_K]$ via a sparse gating module to modulate corrupted expert features and output the high-purity routing sequence E_{route} . **(Fusion and Optimization)** A global–local fusion module performs residual addition between the intact global anchor F_{global} and dimension-aligned local features F_{local} to yield the final identity descriptor F_{final} , trained end-to-end with a joint objective L_{total} comprising cross-entropy loss L_{ce} , triplet loss L_{tri} , and sparsity routing entropy loss L_{ent} . The arrows indicate the direction of data flow; colored token blocks distinguish spatial body-part features and processing branches; $\oplus$ denotes residual addition; and the loss boxes denote the three joint optimization objectives.

This network is a micro Mixture of Experts (MoE) system based on the human physical topological structure. Simultaneously, it dynamically decides the feature confidence of each expert through high-level scene semantics. First, the original image is projected into a Patch token sequence. Next, this sequence enters the semantic part expert partition stage. In this stage, the person token sequence is decoupled into K independent semantic expert representations according to the spatial topological structure. Meanwhile, the scene perception branch extracts the latent environmental context from global features. Then, the Prompt-Guided Dynamic Routing (PGDR) network infers the real-time routing weight for each part expert. Finally, the features enhanced by routing weighting and cross-expert collaboration are fused with global features. This fusion generates the final identity descriptor, which is jointly supervised and optimized by the triplet loss and the cross-entropy loss.

To describe this process more rigorously, the entire inference pipeline is formalized as a composite function. Let $X \in \mathbb{R}^{H \times W \times 3}$ denote the input person image tensor, where H and W are the height and width of the image, respectively. First, the image is mapped into a Patch sequence, denoted as $E_{patch} \in \mathbb{R}^{N \times D}$. Here, N is the number of patches, and D is the embedding dimension:

$$E_{patch} = Flatten(X) \cdot W_{proj} + E_{pos} \quad (1)$$

where $Flatten(X)$ represents flattening the original image along the spatial dimension. W_{proj} is a learnable linear projection matrix used to map the flattened image patches into a D -dimensional feature space. E_{pos} is a learnable position embedding vector used to retain the spatial geometric topology relationship of each patch in the original 2D image. Subsequently, based on the prior knowledge of human physical structure, E_{patch} is vertically divided into K local expert sequences $\{E_k\}_{k=1}^K$. At the same time, the routing module generates a scene prompt vector v_{ctx} based on global statistical information, and calculates the corresponding gating routing weight $\omega \in \mathbb{R}^K$:

$$v_{ctx} = \frac{1}{N} \sum_{i=1}^N E_{patch}^{(i)} \quad (2)$$

$$\omega = \text{Softmax} \left(\frac{\max(0, v_{ctx} W_s + S) W_g + b_g}{\tau} \right) \quad (3)$$

where $W_s \in \mathbb{R}^{D \times d_g}$ and $W_g \in \mathbb{R}^{d_g \times K}$ are learnable weight matrices used for dimensionality reduction and feature mapping in the gating network, respectively. $S \in \mathbb{R}^{1 \times d_g}$ represents the predefined scene prompt prior, $b_g \in \mathbb{R}^{1 \times K}$ represents the learnable bias vector, and τ denotes the temperature scaling factor. The Softmax function ensures that the sum of the finally output routing weights $\omega \in \mathbb{R}^{1 \times K}$ is equal to 1. In Equation (4), the routing sequence $E_{route} \in \mathbb{R}^{K \times D}$ is generated by concatenating the scaled local features along the sequence dimension, where $[\cdot, \cdot]$ represents the spatial concatenation operation. The modulated

local features are then re-concatenated along the sequence dimension to generate a unified routing feature sequence E_{route} :

$$E_{route} = [\omega_1 \cdot E_1, \omega_2 \cdot E_2, \dots, \omega_K \cdot E_K] \quad (4)$$

where $[\cdot, \cdot]$ represents concatenation along the sequence dimension. Finally, this dynamic routing sequence is not directly used for classification. Instead, it is fed into the global–local collaborative fusion module. The network extracts the original macroscopic information without local partition as the global identity anchor F_{global} . Then, it superimposes the dimension-aligned routing features onto this anchor as a residual vector. This process generates the final unified identity descriptor F_{final} for supervised training. Through the above pipeline, the SPDR-Net can autonomously perceive scene changes and dynamically adjust feature attention. In the following sections, the internal structure and mathematical implementation of each module are introduced in detail.

3.2. Semantic Part Expert Construction and Local Feature Refinement

To enable the model to cope with the spatially asymmetric interference challenges in open scenes, the system must possess the ability to decouple local representations with independent semantic attributes from global features. Traditional local feature models mostly adopt a hard horizontal slicing strategy based on physical coordinates. However, this strategy will introduce severe semantic misalignment when the bounding box is unaligned or the human posture changes drastically. Therefore, a local expert construction mechanism based on latent semantic aggregation is proposed. Instead of forcing hard pixel-level segmentation, this mechanism introduces human physiological topology priors to adaptively aggregate tokens with similar semantic affiliations into independent expert paths. This is shown in Figure 3.

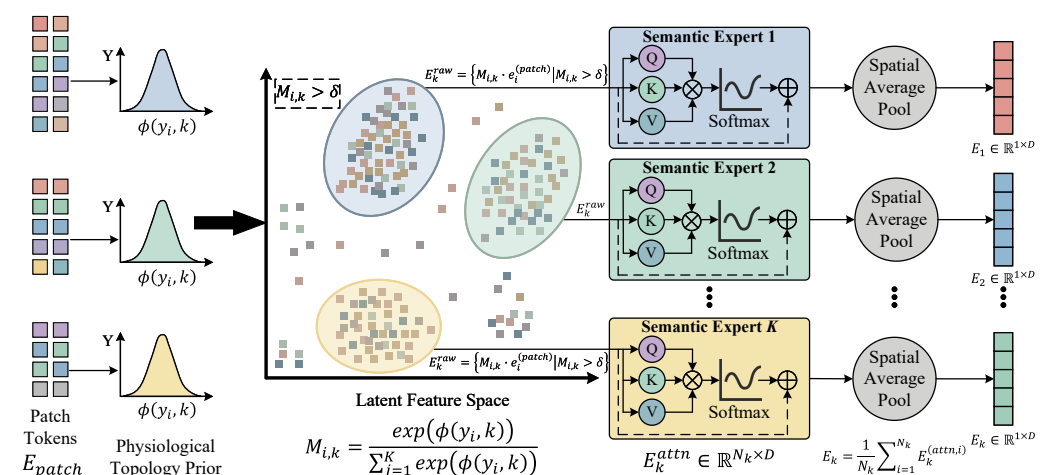


Figure 3. Internal structure of the semantic part expert construction and local feature refinement module. Arrows indicate the processing direction; token colors distinguish patches assigned to different semantic body-part experts; Q, K, and V denote the query, key, and value projections; and the ellipsis denotes intermediate experts or patch tokens omitted only for visual compactness.

First, K latent local semantic centers are uniformly or adaptively defined along the vertical axis. These centers cover the macroscopic spatial distribution of the person from head to foot. The human body structure exhibits a stable topology in the vertical dimension, whereas its horizontal distribution is highly sensitive to camera angles and walking directions. Therefore, we exclusively define the semantic centers along the vertical axis. Given N Patch tokens, the spatial center coordinate of each token i is denoted as (x_i, y_i) .

To achieve semantic alignment, the correlation factor $M_{(i,k)}$ between token i and the k -th local expert is calculated:

$$M_{i,k} = \frac{\exp(\phi(y_i, k))}{\sum_{j=1}^K \exp(\phi(y_i, j))} \quad (5)$$

where $\phi(\cdot)$ is a spatial response function based on the Gaussian distribution. Its center points are sequentially and equidistantly preset along the vertical central axis of the human body. This soft assignment mechanism ensures that even when the person posture deviates slightly from the center, related patches can still be correctly classified into the corresponding semantic expert paths.

In Equation (5), the spatial response function $\phi(y_i, k)$ based on the Gaussian distribution is mathematically formulated as

$$\phi(y_i, k) = -\frac{(y_i - \mu_k)^2}{2\sigma^2} \quad (6)$$

where y_i represents the vertical coordinate of the i -th patch token, μ_k represents the preset vertical center coordinate of the k -th semantic part, and σ represents the standard deviation controlling the spatial response range. This soft assignment mechanism ensures that even when the person posture deviates slightly from the center, related patches can still be correctly classified into the corresponding semantic expert paths.

Subsequently, through weighted aggregation of the tokens, K initial local expert features $E_k^{raw} \in \mathbb{R}^{N_k \times D}$:

$$E_k^{raw} = \left\{ M_{i,k} \cdot e_i^{(patch)} \mid M_{i,k} > \delta \right\} \quad (7)$$

where δ is a sparse threshold for filtering background noise. $e_i^{(patch)}$ represents the D -dimensional feature vector obtained after the i -th patch passes through the Transformer linear embedding and position encoding. This process decouples the lengthy global sequence into K candidate feature clusters with independent semantic directions. Although this aggregation process achieves initial semantic clustering, the feature sequence within each expert path still contains redundant noise due to the influence of background clutter (e.g., trees, persons) in open scenes. To extract high-purity identity discriminative signals within each expert, an independent self-attention mechanism is introduced for each local expert to perform secondary local feature refinement and suppress background interference.

For the k -th local expert E_k^{raw} , we first generate independent Query (Q_k), Key (K_k), and Value (V_k) matrices through linear mapping:

$$Q_k = E_k^{raw} \times W_q^{(k)} \quad (8)$$

$$K_k = E_k^{raw} \times W_k^{(k)} \quad (9)$$

$$V_k = E_k^{raw} \times W_v^{(k)} \quad (10)$$

where $W_q^{(k)}, W_k^{(k)}, W_v^{(k)} \in \mathbb{R}^{D \times D}$ are linear projection weights unique to the k -th part expert. These weights are optimized independently and are not shared with other experts. Subsequently, the similarity attention map within this expert is calculated and feature-weighted denoising is performed:

$$E_k^{attn} = \text{Softmax} \left(\frac{Q_k \times K_k^T}{\sqrt{d_h}} \right) V_k + E_k^{raw} \quad (11)$$

where d_h is the scaling factor of the attention mechanism. To clearly define the tensor dimensions in Equation (11), the Query, Key, and Value matrices are represented as Q_k, K_k, V_k , respectively. The learnable linear projection weights are unique as $W_q^{(k)}, W_k^{(k)}, W_v^{(k)} \in \mathbb{R}^{D \times D}$. The Softmax function operates row-wise on the computed similarity matrix of shape $\mathbb{R}^{N_k \times N_k}$ to model the contextual correlations. The output $E_k^{attn} \in \mathbb{R}^{N_k \times D}$ represents the pure local feature after part-level attention denoising. We clarify that the implicit denoising capability of the self-attention mechanism within each semantic expert is achieved through spatial weight redistribution. In open-world scenes, a cropped pedestrian bounding box often contains redundant background clutter, such as trees, vehicles, or background pedestrians. Since these background patches lack semantic correlation with the target pedestrian's body parts, the similarity calculation in Equation (11) assigns extremely low attention weights to these non-salient regions. Conversely, the discriminative foreground patches belonging to the pedestrian body receive high attention weights. By performing feature-weighted summation, this self-attention block naturally filters out background noise and concentrates the feature energy on the clean body structures, thereby achieving implicit semantic refinement within each expert's subspace. To unify the feature dimensions to match the weight calculation of the subsequent routing network, spatial average pooling is performed on the denoised features E_k^{attn} along the sequence dimension. This operation compresses it into a single expert descriptor:

$$E_k = \frac{1}{N_k} \sum_{i=1}^{N_k} E_k^{(attn,i)} \quad (12)$$

Through the above pipeline, K highly refined local expert feature vectors are finally obtained. $\{E_1, E_2, \dots, E_K\}$, where each vector dimension is $E_k \in \mathbb{R}^{1 \times D}$. These expert features not only preserve the local fine-grained differences of the person but also achieve initial fault tolerance for posture changes through the latent semantic aggregation mechanism. However, when facing extreme spatially asymmetric interference (e.g., a certain part is completely occluded), the reliability of these expert features is not equal. Therefore, the model needs an adaptive decision to assign transmission weights to each expert feature.

3.3. Prompt-Guided Dynamic Routing

After extracting the local expert features E_k , the model obtains a reliable feature foundation. However, the model still needs to address the challenge of spatially asymmetric interference in the open world. Traditional models usually adopt a static concatenation fusion strategy. Once a certain body part suffers from severe occlusion or illumination variation, its generated noise features will inevitably affect the overall identity representation. If each local expert is regarded as an independent information transmission path, an optimal strategy for mitigating asymmetric noise is to establish a control center. This center dynamically controls each information transmission path through a dynamic routing decision. This router can act as a soft logical switch. Based on the overall environmental quality of the current image, it adaptively determines the information bandwidth (i.e., weight) of each local feature path.

Based on the above problems, the Prompt-Guided Dynamic Routing (PGDR) module is proposed. This module achieves the dynamic allocation of local expert weights by introducing high-level scene semantics. Before executing the environment-driven path decision, the model first needs to perceive the current macroscopic environmental state (e.g., whether it is in low illumination, or whether there is large-area physical occlusion). In Section 3.1, the system has extracted the environmental context representation $v_{ctx} \in \mathbb{R}^{1 \times D}$ through Global Average Pooling. However, the original visual context often contains redundant information and lacks the high-level semantic discriminative power to guide complex

decisions. Therefore, PGDR introduces a learnable scene prompt prior $S \in \mathbb{R}^{1 \times d_g}$ (where $d_g \ll D$ is the compressed decision space dimension). This prompt vector is continuously updated through backpropagation during training. It remembers the universal semantic distribution patterns across various domain scenes. Inside the gating network, joint spatial alignment is first performed between the high-dimensional visual context v_{ctx} and the semantic prior S to generate a compact fusion decision basis U_{joint} :

$$U_{joint} = v_{ctx} \cdot W_s + S \quad (13)$$

where $W_s \in \mathbb{R}^{D \times d_g}$ is the spatial dimensionality reduction mapping matrix. This operation strips away useless low-level pixel noise and compresses the decision signal into a pure high-order semantic space. A natural question may arise regarding the self-referencing relationship in Equation (12), where the global context v_{ctx} and the local experts $\{E_k\}$ are derived from the same initial patch tokens. We clarify that this design represents a mathematically sound and highly efficient macro-to-micro feedback loop. Although they share the same source, v_{ctx} extracts the low-frequency global style of the image (e.g., overall illumination, weather conditions, and background color temperature) by average pooling the entire sequence. In contrast, the local experts $\{E_k\}$ preserve the high-frequency localized physical topology of the human body. Using the global pooled representation of a system to gate or modulate its own localized components is a well-established paradigm in computer vision, such as in Squeeze-and-Excitation (SE) blocks or spatial-channel attention networks. This top-down feedback mechanism allows the macroscopic scene semantics to guide the trust allocation of microscopic details with zero additional parameter overhead.

After obtaining the fusion basis, the network needs to generate the corresponding unnormalized routing probabilities for the K part experts. To endow the network with the ability to strongly suppress a specific feature path under extreme interference, the sparse non-linear characteristic of the *ReLU* function is utilized to perform truncation activation on U_{joint} :

$$h_{gating} = \max(0, U_{joint}) \quad (14)$$

$$L_{route} = h_{gating} \times W_g + b_g \quad (15)$$

where h_{gating} represents the gating feature vector after sparse truncation activation. $W_g \in \mathbb{R}^{d_g \times K}$ and $b_g \in \mathbb{R}^{1 \times K}$ are the linear projection weight matrix and bias term, respectively. They map the decision spatial features to the probability of each part expert. $L_{route} = [l_1, l_2, \dots, l_K] \in \mathbb{R}^{1 \times K}$ is the unnormalized routing probability vector output by the network. The magnitude of each element inside this vector represents the initial evaluation score of the model regarding the reliability of the corresponding part expert.

Finally, to convert the discrete probability scores into routing weights $\omega \in \mathbb{R}^{1 \times K}$ that satisfy the constraints of probability theory, the *Softmax* function is applied to normalize L_{route} . In addition, to further sharpen the weight distribution (i.e., enlarge the weight gap between high-confidence and low-confidence parts), a temperature scaling factor τ is introduced:

$$\omega_k = \frac{\exp(l_k/\tau)}{\sum_{j=1}^K \exp(l_j/\tau)}, \quad k \in \{1, 2, \dots, K\} \quad (16)$$

where l_k and l_j represent the k -th and j -th elements in the routing probability vector L_{route} , respectively. τ is the temperature variable controlling the smoothness of the probability distribution. In all experiments, τ is fixed at 1.0 during both training and inference. $\omega_k \in (0, 1)$ is the final dynamic routing weight assigned to the k -th part expert after normalization.

We also address the mathematical properties of the softmax function in Equation (16). Theoretically, the output of a standard softmax function is strictly positive, meaning that a strongly suppressed expert path still retains a non-zero weight. In our formulation, the ReLU operation in Equation (14) is applied to the intermediate gating representation U_{joint} , producing $h_{gating} = \max(0, U_{joint})$; it does not directly truncate the routing logits. The logits $L_{route} = h_{gating} \times W_g + b_g$ in Equation (15) are subsequently generated by a linear projection and may still take negative values. Therefore, the routing sparsity should be understood as originating from the non-negative hidden gating representation, the entropy regularization, and the temperature-controlled Softmax normalization, rather than from forcing negative logits to zero. Under extreme local interference, this design drives the gating network to generate highly polarized but strictly positive routing weights. For instance, the weight assigned to a severely corrupted leg expert can be suppressed to a very small scale (e.g., $< 10^{-4}$). This value is not numerically equivalent to zero in floating-point arithmetic; instead, it acts as a soft attenuation factor that proportionally reduces the contribution of the corrupted feature path without mathematically setting its weight to zero. Ultimately, this weight is used as a scalar factor through an element-wise broadcasting mechanism. It performs dynamic modulation on the input expert features to generate the final routing sequence:

$$E_{route} = \text{Concat}\left(\{\omega_k \times E_k\}_{k=1}^K\right) \quad (17)$$

Here, we provide a crucial mathematical clarification regarding the combination of the softmax normalization and the concatenation operation. In conventional Mixture of Experts (MoE) networks, experts typically share a unified output space. Thus, their features are merged through a weighted sum. However, in our SPDR-Net, each semantic expert represents a physically distinct spatial region of the human body (e.g., head, torso, and lower limbs). These spatial components reside in complementary and non-overlapping semantic subspaces. Applying a traditional weighted sum would physically mix these distinct features, causing a catastrophic loss of spatial topology. Therefore, we deliberately adopt the concatenation operation in Equation (17) to preserve the integrity of the vertical body structure. Under this formulation, the sum-to-one constraint should not be interpreted as providing indispensable absolute feature magnitudes. In a clean scene, when all local experts are reliable, the routing distribution naturally tends to a nearly uniform state, i.e., $\omega_k \approx 1/K$. This operation indeed introduces an almost uniform scaling over all local expert slots, and this global scale can be absorbed by the subsequent projection matrix W_f in the global-local fusion module. Therefore, the effective decision information carried by the routing weights lies in their relative proportions rather than their absolute magnitudes. (1) Eliminating scale degree of freedom: The role of the sum-to-one constraint is to remove the scale degree of freedom in the gating branch and map the routing vector onto a probability simplex, so that the network expresses relative expert reliability instead of arbitrary feature amplification or attenuation. (2) Automatic gain control: From an energy-stabilization perspective, this normalization also acts as an automatic gain control mechanism. If independent gates such as Sigmoid outputs were used without normalization, multiple clean experts could simultaneously receive weights close to one, whereas an occluded image might retain only a few active experts. In that case, the total energy of the local branch would vary with the number of visible parts, introducing a feature-norm fluctuation unrelated to identity discrimination. By enforcing $\sum_{k=1}^K \omega_k = 1$, the gating branch cannot arbitrarily inflate or shrink the overall local feature scale; it can only redistribute trust among experts. (3) Well-defined Shannon entropy: this scale-normalized formulation makes the routing entropy regularization mathematically valid and well-defined, ensuring that routing decisions reflect relative reliability rather than uncontrolled norm changes.

The above algorithm realizes the process of dynamic routing decision-making. For example, when the model faces an extreme scene where the person's lower limbs are severely occluded, the weight distribution obtained after non-linear reasoning by the sparse gating network will show a polarization tendency (e.g., $\omega = [0.25, 0.25, 0.25, 0.15, 0.05, 0.05]$). Through this algorithm, the leg feature paths polluted by noise are assigned much lower transmission weights. Meanwhile, the clear head and torso paths are endowed with higher information transmission bandwidths. This strongly supports the robustness of E_{route} in complex open scenes. The detailed workflow is illustrated in Figure 4.

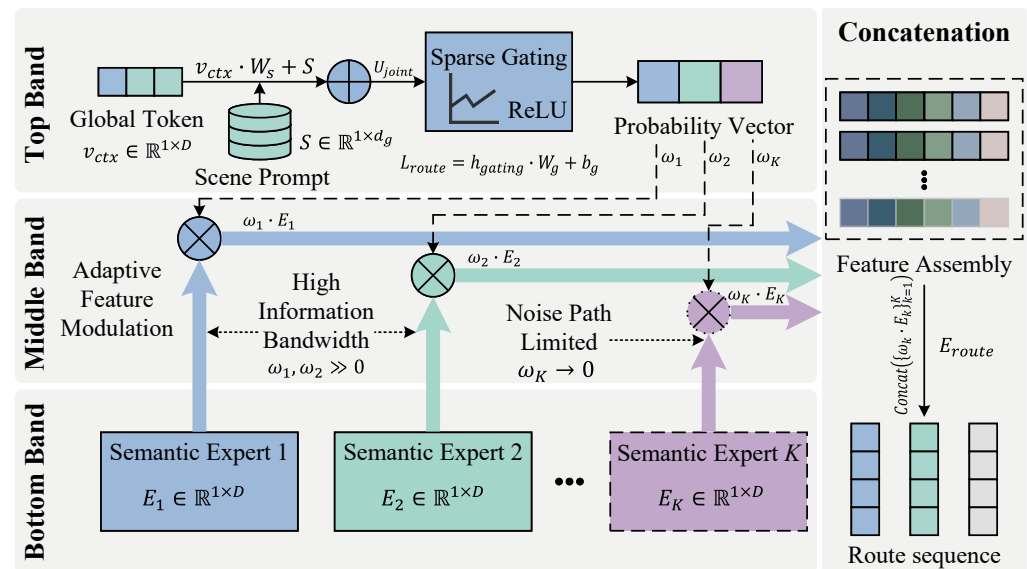


Figure 4. Internal structure of the prompt-guided dynamic gating routing module. Arrows indicate information flow; different colors distinguish body-part expert paths and their routed features; solid and attenuated paths indicate high- and low-weight routes, respectively; $\oplus$ and $\otimes$ denote addition and feature modulation; and the ellipsis denotes intermediate experts omitted for visual compactness.

3.4. Global–Local Feature Fusion

The local feature sequence E_{route} modulated by the scene-prompt dynamic routing demonstrates superior noise robustness and detail refinement ability under extreme interference. However, relying solely on isolated local sequences for identity retrieval tends to fragment the inherent macroscopic topological structure of the human body (e.g., the color matching continuity of the overall clothing). This fragmentation can lead to misjudgments when facing highly similar negative samples (e.g., different persons wearing the same style/color of clothing).

To address this limitation of local features, this paper designs a global–local feature collaborative fusion module. This module aims to bridge the semantic gap between the global macroscopic identity anchor and the local detailed features for the model. Consequently, it generates a unified identity descriptor possessing both topological structure integrity and fine-grained discriminative power, as illustrated in Figure 5. First, the overall environmental and identity representation that has not been destroyed by the local partitioning operation needs to be acquired. Given the global Patch feature sequence $E_{global} \in \mathbb{R}^{N \times D}$ containing all spatial information output by the deep network (where N is the total number of Patch tokens and D is the feature dimension), a Global Average Pooling (GAP) operation is executed along the spatial dimension (i.e., the sequence direction):

$$F_{global} = \left(\frac{1}{N} \sum_{i=1}^N E_{global}^{(i)} \right) \cdot W_{proj_g} + b_{proj_g} \quad (18)$$

where $E_{global}^{(i)} \in \mathbb{R}^{1 \times D}$ represents the i -th row of the global patch feature sequence E_{global} . Furthermore, $W_{proj_g} \in \mathbb{R}^{D \times C}$ and $b_{proj_g} \in \mathbb{R}^{1 \times C}$ are the learnable projection weight matrix and bias vector, respectively. Through this linear projection, the pooled feature vector of shape $\mathbb{R}^{1 \times D}$ is mapped into the target C -dimensional embedding subspace, constituting the macroscopic baseline $F_{global} \in \mathbb{R}^{1 \times C}$ for person identity discrimination. Simultaneously, the dynamic routing sequence E_{route} output in Section 3.3 contains the discrete features of K experts, with a dimension of $\mathbb{R}^{K \times D}$. To enable E_{route} to effectively interact with the global anchor in the same metric space, it needs to be geometrically flattened and spatially projected:

$$F_{local} = Flatten(E_{route}) \cdot W_f + b_f \quad (19)$$

where the $Flatten(\cdot)$ operation concatenates the features of the K local experts along the channel dimension, converting them into a 1D tensor of $\mathbb{R}^{1 \times K \cdot D}$. Subsequently, through a learnable linear projection matrix $W_f \in \mathbb{R}^{(K \cdot D) \times C}$ and a bias term $b_f \in \mathbb{R}^{1 \times C}$, the high-dimensional concatenated local features are precisely mapped into the same C -dimensional subspace as F_{global} . The generated F_{local} contains the high-purity local detailed features filtered by the routing weights.

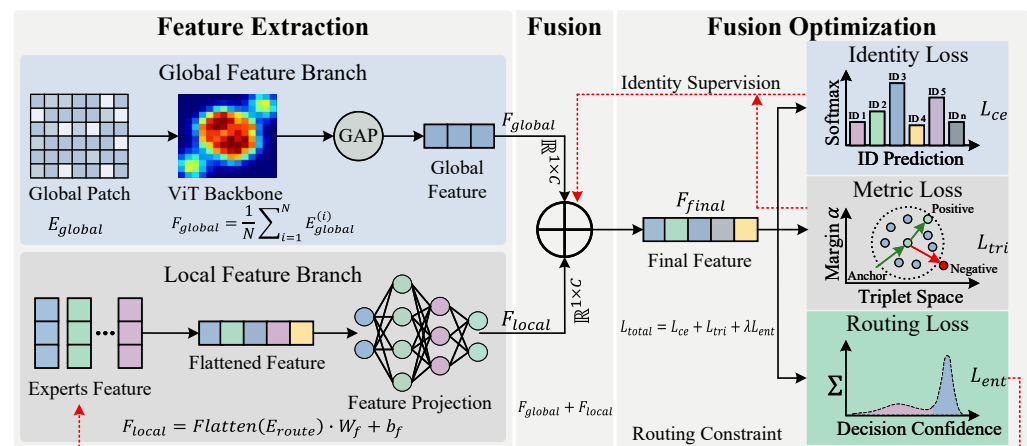


Figure 5. Internal structure of the global–local feature collaborative fusion module. Arrows indicate the feature and optimization flows; blue and green blocks denote global and local feature branches; $\oplus$ denotes residual fusion; the red brackets group the identity, metric, and routing optimization objectives; and the ellipsis in the identity classifier denotes intermediate classes omitted for visual compactness.

After completing the spatial alignment, the network does not adopt the traditional channel concatenation strategy. Instead, it introduces an element-wise addition residual fusion mechanism:

$$F_{final} = F_{global} + F_{local} \quad (20)$$

where $F_{final} \in \mathbb{R}^{1 \times C}$. In this element wise addition residual fusion mechanism, F_{global} is treated as the baseline of the identity representation, while F_{local} is regarded as the residual compensation vector targeting detailed features. This mechanism not only avoids the redundant expansion of feature dimensions but also forces the network to learn those subtle differences that the global module cannot capture during backpropagation.

3.5. Optimization Objective

After passing through the global–local feature collaborative fusion module, the network finally outputs a unified identity descriptor F_{final} . This descriptor possesses both macroscopic environmental robustness and detailed discriminative power. To drive the end-to-end optimization of the SPDR-Net and ensure the correctness of the dynamic routing decision, a multi-task joint objective loss function is designed. This function includes identity classification, metric learning, and sparse routing penalty.

First, the generated feature vector $F_{final} \in \mathbb{R}^{1 \times C}$ is fed into a classifier. This classifier consists of a linear fully connected layer and a *Softmax* activation function. It predicts the probability distribution p of the person belonging to each known identity category. In this process, the standard cross-entropy loss is used for optimization:

$$L_{ce} = - \sum_{j=1}^M y_j \log(p_j) \quad (21)$$

where M is the total number of identity categories in the training set. y_j is the one-hot encoding indicator function of the true label. This loss aims to optimize the distribution of features and ensure the linear separability between different identity features. Simultaneously, a hard-mining triplet marginal loss is introduced. This loss aims to further pull closer intra-class samples and push away inter-class samples in the feature metric space. Within a training batch, the feature of the extracted anchor sample is denoted as F_a . The positive sample feature F_p farthest from it within the batch is searched, and the closest negative sample feature F_n is also identified. The formal definition of the triplet loss is as follows:

$$L_{tri} = \max(0, \|F_a - F_p\|_2 - \|F_a - F_n\|_2 + \alpha) \quad (22)$$

where $\|\cdot\|_2$ represents the L_2 norm of the Euclidean distance. α is a preset distance margin threshold. This margin is used to force the cohesion of intra-class samples to be strictly better than the distance between inter-class samples.

Although the aforementioned two conventional losses guarantee the basic effectiveness of the identity representation, these metrics cannot directly constrain the weight allocation inside the PGDR module. Under the condition of lacking explicit supervision, the *Softmax* gating easily falls into the dilemma of “trivial decision”. That is, no matter what kind of scene it faces, it always outputs a roughly uniform routing weight. To address this problem and force the gating network to make decisive decisions under scene interference, A sparsity routing entropy loss (L_{ent}) based on information theory is proposed. For training data with a batch size of B , the routing weight distribution of the i -th sample on K local experts is denoted as $\omega^{(i)} = [\omega_1^{(i)}, \dots, \omega_K^{(i)}]$. The Shannon information entropy of this probability distribution is calculated and minimized:

$$L_{ent} = \frac{1}{B} \sum_{i=1}^B \sum_{k=1}^K -\omega_k^{(i)} \log(\omega_k^{(i)} + \epsilon) \quad (23)$$

where $\epsilon = 10^{-8}$ is a small constant to prevent numerical underflow of the logarithmic function. By minimizing L_{ent} , the noise features of the affected parts can be effectively suppressed. When the network’s confidence in each part is uncertain (i.e., the weights tend to be uniform), the information entropy reaches its maximum. Conversely, when the network is certain to keep one part and strongly suppress another, the Softmax-normalized weights remain strictly positive but can become highly polarized (e.g., $[0.8, 0.1, 0.0997, 10^{-4}, 10^{-4}, 10^{-4}]$), and the information entropy decreases substantially rather than relying on exact zero-valued gates.

Based on the above comprehensive analysis, the final joint training loss function L_{total} of the SPDR-Net consists of three weighted parts:

$$L_{total} = L_{ce} + L_{tri} + \lambda L_{ent} \quad (24)$$

where λ is a hyperparameter balancing the strength of the routing entropy penalty. Through the backpropagation of this joint loss function, the network ensures global identity discriminative power. Meanwhile, it adaptively adjusts the interference to local asymmetry.

Conceptually, minimizing the unsupervised routing entropy L_{ent} alone might lead to a trivial collapse. Without additional constraints, the network might always route the weights to one single fixed expert (e.g., always choosing the head expert with a limiting distribution approaching $[1, 0, 0]$). To prevent this routing collapse, our SPDR-Net is trained under a multi-task joint optimization objective, as defined in Equation (24). The identity classification loss L_{ce} and the hard-mining triplet loss L_{tri} act as powerful regularizers against this trivial solution. Under normal or clean surveillance environments, the network must preserve the comprehensive features of all body parts to successfully classify and match diverse pedestrian identities. If the router collapses to a single expert, the model loses the discriminative details of other parts. This loss causes a catastrophic spike in both L_{ce} and L_{tri} . Consequently, the joint optimization establishes a competitive game: the identity losses encourage the model to maintain balanced and holistic features in clean scenes, whereas L_{ent} continuously encourages lower-entropy routing over all samples. Therefore, the conditional selectivity of polarized routing under local asymmetric interference should be interpreted as an emergent behavior from the balance among L_{ce} , L_{tri} , and L_{ent} , rather than as a hard conditional mechanism that is activated only after noise is explicitly detected.

3.6. Overall Implementation Workflow

The overall workflow of the SPDR-Net is summarized in Algorithm 1.

Algorithm 1 Forward inference and optimization of SPDR-Net

Require: Training batch images $\mathcal{I} \in \mathbb{R}^{B \times H \times W \times 3}$, Labels Y , ViT backbone Ψ , Prompt S .

Ensure: Optimized network parameters and Prompt S .

```

1: for each training iteration do
2:   % Step 1: Feature Embedding & Local Feature Refinement
3:   Extract  $E_{patch}, E_{global} \in \mathbb{R}^{B \times N \times D}$  via Patch Embedding and  $\Psi$ .
4:   Partition  $E_{patch}$  spatially into  $K$  expert sequences  $\{E_k^{raw}\}_{k=1}^K$ .
5:   for  $k = 1$  to  $K$  do                                     ▷ Parallelizable across experts
6:      $E_k \leftarrow \text{AvgPool}\left(\text{Softmax}\left(\frac{Q_k K_k^T}{\sqrt{d_h}}\right) V_k + E_k^{raw}\right)$ 
7:   end for
8:   % Step 2: Scene-Prompt-Guided Dynamic Routing
9:    $v_{ctx} \leftarrow \text{AvgPool}(E_{patch})$                                ▷ Extract visual context
10:  Routing weights  $\omega \leftarrow \text{Softmax}\left(\left(\max(0, v_{ctx} W_s + S) W_g + b_g\right) / \tau\right)$ 
11:   $E_{route} \leftarrow \text{Concat}_{k=1}^K (\omega_k \cdot E_k)$                                ▷ Dynamic modulation
12:  % Step 3: Global-Local Feature Fusion
13:   $F_{global} \leftarrow \text{AvgPool}(E_{global})$ 
14:   $F_{final} \leftarrow F_{global} + (\text{Flatten}(E_{route}) W_f + b_f)$            ▷ Residual fusion
15:  % Step 4: Multi-Task Joint Optimization
16:  Compute ID Loss  $L_{ce}$  and Triplet Loss  $L_{tri}$  using batch  $\{F_{final}, Y\}$ .
17:  Compute Sparsity Entropy Loss:  $L_{ent} \leftarrow \frac{1}{B} \sum_{i,k} -\omega_{i,k} \log(\omega_{i,k} + \epsilon)$ 
18:  Update parameters and  $S$  via backpropagation of  $L_{total} = L_{ce} + L_{tri} + \lambda L_{ent}$ .
19: end for

```

The complete forward inference and backward optimization process of this framework can be refined into the following four steps: (1) Feature embedding and local feature refinement stage. The input image is processed into a Patch grid. Then, it is evenly divided into K original local expert sequences using the human spatial topology prior. Subsequently, each sequence performs denoising and refinement internally through the self-attention mechanism. These sequences are then compressed into independent high-quality expert descriptors by spatial pooling. (2) Scene-prompt-guided dynamic routing stage. The network extracts the macroscopic visual context of the current image. It then performs joint spatial alignment with the learnable scene prompt prior. Next, the sparse gating network infers the routing weight distribution against spatially asymmetric interference. Finally, the network adaptively and dynamically modulates each part expert feature. (3) Global–local feature collaborative fusion stage. The network performs element-wise residual addition. It adds the global macroscopic identity anchor (containing the original topological structure) to the dimensionality-aligned local routing sequence. Thus, it outputs the final unified identity descriptor possessing both macroscopic environmental robustness and detailed discriminative power. (4) Multi-task joint constraint and end-to-end optimization stage. Based on the final identity descriptor, the system calculates the identity classification cross-entropy loss and the hard-mining triplet loss in parallel. Simultaneously, it uses the sparsity routing entropy penalty loss to constrain the confidence of the gating decisions. Finally, the three losses are weighted and summed to form a joint objective function. The dynamic update of the model parameters is realized through backpropagation.

4. Experiments

4.1. Datasets and Evaluation Protocol

To comprehensively verify the scene adaptability and anti-interference robustness of the proposed SPDR-Net in open-world environments, this paper conducts detailed evaluations on seven mainstream public person ReID datasets and the newly constructed SD-ReID dataset. These datasets cover various challenges, including normal, occlusion, cross-modality (infrared), and clothing-changing scenes. Based on their contents and properties, these datasets are categorized as follows: normal ReID datasets (Market1501 [37] and MSMT17 [38]), occluded ReID dataset (Occ-Duke [39]), low-resolution ReID dataset (MLR-CUHK03 [40]), infrared ReID dataset (SYSU-mm01 [41]), clothing-changing ReID datasets (Celeb-ReID [42] and PRCC [43]), and a newly constructed street scene ReID dataset (Street Deployment ReID, SD-ReID [44]). In all experiments, the official training and evaluation protocols of all datasets are followed. The mean Average Precision (mAP) and Rank-1 accuracy are primarily reported. Introductions to the above datasets are as follows.

(1) Market1501 and Msmt17 datasets. The Market1501 dataset includes 32,668 images of 1501 labeled persons, captured from six camera perspectives. There are 751 identities in the training set and 750 identities in the test set. The dataset is divided into three parts, namely train, gallery and query, with the latter two constituting the testing set.

The Msmt17 dataset employs a network of 15 cameras on campus, including 12 outdoor cameras and 3 indoor cameras, with a total of 126,441 bounding frames for 4101 persons. Among them, the training set contains 1041 persons with a total of 32,621 bounding boxes, while the test set includes 3060 persons with a total of 93,820 bounding boxes.

(2) Occ-Duke dataset. The Occ-Duke dataset is a specialized subset derived from the DukeMTMC-reID dataset. It specifically selects images with severe occlusions to evaluate model robustness against physical occlusions. A total of 1404 people and 36,411 images are collected from 8 different cameras. Among them, 702 people are used as the training set and 702 people as the test set.

(3) MLR-CUHK03 dataset. MLR-cuhk03 is a variant of the CUHK03 dataset. It creates low-resolution query images by downsampling CUHK03 query images. Half of the training images of MLR-cuhk03 are downsampled versions of CUHK03. The MLR-CUHK03 dataset contains a total of 965 low-resolution query images, 977 high-resolution gallery images, and 26,251 training images.

(4) SYSU-mm01 dataset. The SYSU-MM01 dataset is an image dataset for visible light-infrared person re-identification. The images are collected by 4 RGB cameras (for bright scenes) and 2 infrared cameras (for dim scenes). In total, these 6 cameras capture 30,071 RGB images and 15,792 infrared images of 491 persons.

(5) Celeb-reID and PRCC dataset. The images in the Celeb-reID dataset are produced using street snapshots of celebrities obtained from the Internet. There are a total of 1052 identities and 34,186 images, making it the largest long-term costume ReID dataset to date.

The PRCC dataset consists of 33,698 images from 221 identities. Subjects captured by Cameras A and B wear identical clothing, whereas clothing changes are introduced in Camera C. However, in Camera C, the person wears different clothes. The images are taken on different dates.

(6) SD-ReID Dataset. Although existing academic datasets have greatly promoted the development of ReID technology, most of them are collected in relatively controlled campus or laboratory environments. Their viewpoints are relatively single, and their backgrounds are relatively clean. In real open-world video surveillance (e.g., city intersections, commercial person streets), person features often face extremely complex domain shift phenomena simultaneously. To ensure that real person images reflect spatially asymmetric interference, and to provide a benchmark verification library for actual ReID engineering deployment, we construct the SD-ReID dataset.

The original images of the SD-ReID dataset originate from publicly available high-definition surveillance video streams of real sidewalks and complex blocks on the Internet. The entire dataset is captured by 6 cameras at different heights and viewpoints. It covers extreme viewpoints such as flat, top-down, and side views. Compared with traditional datasets, the SD-ReID dataset possesses the following three significant open-world challenge characteristics: (1) Frequent partial dynamic occlusion: In real streets, persons are frequently occluded by carried items, vehicles, and other persons. This occlusion is often asymmetric (e.g., only occluding the lower body or torso). This phenomenon frequently leads local features to adversely impact the overall person recognition. (2) Drastic scale and resolution variance: Because surveillance cameras have a large depth of field, the effective pixel resolution of pedestrians changes significantly as they walk toward the camera. Consequently, this variance often causes traditional models to generate geometric misalignment when separating person features. (3) High-similarity distractors and illumination changes: The SD-ReID dataset contains many person images with similar clothing and local illumination changes. This phenomenon easily causes person local features to be affected, thereby misidentifying the person's identity.

To standardize person images and increase dataset compatibility, we performed strict bounding box annotation and cropping on the collected original video frames. All person images in the SD-ReID dataset are uniformly scaled to a fixed resolution of 64×128 pixels. For images that do not meet this size, a zero-padding strategy is adopted to match the image dimensions. In addition, SD-ReID strictly follows the standard naming system of classic ReID datasets like Market-1501. The file name of each image precisely encodes the identity and spatiotemporal information of the target. Its file naming format is defined as follows:

$$\text{ID_cXsY_ZZZZZZ} . \text{jpg} \quad (25)$$

where ID (e.g., 0045) represents the unique identity label of the person, cX (e.g., c3) indicates the X-th camera index capturing the image (ranging from c1 to c6), sY represents the specific video clip sequence number under that camera, and ZZZZZZ is the frame number of the image in the original video. This standard format not only supports identity-based metric learning but also reserves expansion space for future cross-camera tracking research based on spatiotemporal topology. The samples of the SD-ReID dataset are shown in Figure 6.

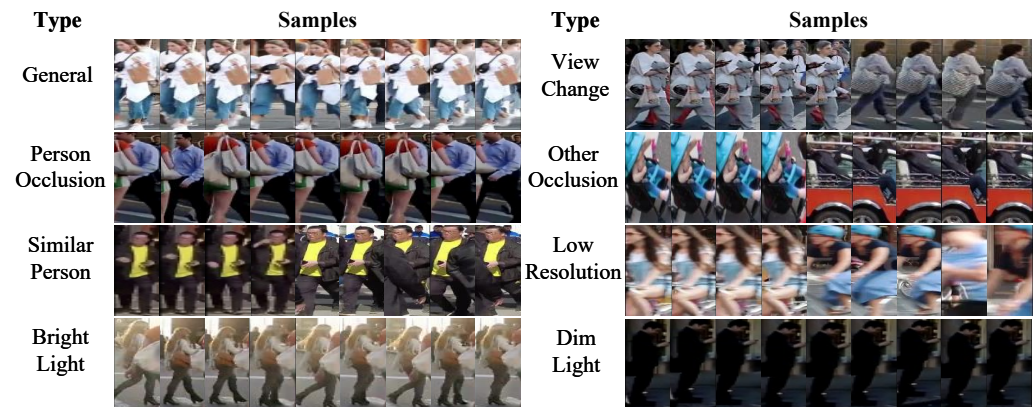


Figure 6. Visualization of samples from the proposed SD-ReID dataset.

After strict screening and cleaning, the SD-ReID dataset contains a total of 10,429 high-quality annotated images of 147 independent identities. To ensure the objectivity of the model evaluation, we divide the involved identity set into non-overlapping training and testing sets according to the principle of zero-shot cross-validation:

- (1) Training Set: Contains 8374 images of 117 identities. This subset is used to guide the model to learn scene prompt priors and optimize the routing decision boundaries of the gating network.
- (2) Contains the remaining 2055 images of 30 identities. During the testing phase, the SD-ReID dataset is further divided into a query set (Query) and a gallery set (Gallery). Because the testing set identities are completely invisible to the model, it can objectively reflect the generalization retrieval ability of the model in a real surveillance system.

To fully specify the SD-ReID protocol, we provide the exact train, query, and gallery splits used in our experiments. Under the standard evaluation protocol, the training set contains 8374 images of 117 identities. The remaining 2055 images of 30 identities constitute the testing set, which is further divided into a query set of 620 images and a gallery set of 1435 images.

Furthermore, to address the potential overfitting issue on this dataset, we also introduce a highly challenging Cross-Camera and Cross-Time (CCCT) evaluation protocol. Under this CCCT protocol, the query set (512 images) and the gallery set (1543 images) are separated both spatially (different camera views) and temporally (different time intervals of the day). This setup reduces short-term frame redundancy and provides a much more rigorous benchmark to evaluate the generalization limit of ReID models in practical security systems.

The SD-ReID dataset uses Rank-1 and mAP indicators to evaluate model performance. The evaluation is conducted in single-query mode. It is widely acknowledged that the accessibility of data and code is vital for scientific progress. To address reproducibility concerns, we have organized the SD-ReID dataset, the source code of SPDR-Net, and the pretrained model weights. These resources are publicly available at GitHub (version 1.0) as the SPDR-Net software release corresponding to this article upon acceptance of this

paper [44]. This open-source availability allows other researchers to reproduce our results and benchmark their methods on this newly constructed dataset.

4.2. Implementation Details

To ensure the fairness and reproducibility of the experiments, all model training and testing in this paper are conducted under a unified software and hardware environment. Specifically, the experimental platform runs on the Linux operating system (Ubuntu 22.04). It is equipped with a single NVIDIA RTX 4090D (24GB VRAM) and an Intel Xeon Platinum 8481C (16 vCPU) processor. The deep learning framework uniformly adopts PyTorch 2.3.0 and CUDA 12.1. The overall training and evaluation pipeline of SPDR-Net follows a strict standardization protocol. The specific implementation details mainly cover the following three aspects.

(1) Model initialization and data preprocessing. The Vision Transformer backbone network (ViT-Base) of SPDR-Net loads pretrained weights on ImageNet-1K in the early stage of training. During the data input stage, the original images in all datasets are uniformly resized to 384×128 pixels using bilinear interpolation. To cope with the ubiquitous environmental interference in the open world and suppress overfitting, multiple data augmentation strategies are introduced. These specifically include: random horizontal flipping with a probability of 0.5, random erasing with a probability of 0.5, and a random dropout mechanism (the dropout rate is set to 0.1).

(2) Training strategy and hyperparameter settings. During the training process, the standard PK sampling strategy is adopted to construct the training batch. That is, the size of each batch is set to $B = 64$. This batch contains $P = 16$ randomly selected different identities, and $K = 4$ images are selected for each identity. This ensures the effectiveness of hard mining in the triplet loss. The number of multi-thread data loaders is set to 16. The model uses the Adam optimizer for parameter updating. The base learning rate is initialized to 3×10^{-4} , and a cosine annealing decay strategy is adopted with a minimum learning rate of 0. The entire training process lasts for 60 epochs.

(3) Joint loss function parameter configuration. The total objective function of the model consists of the cross-entropy loss L_{ce} , the triplet loss L_{tri} , and the sparsity routing entropy loss L_{ent} . In specific hyperparameter settings, the triplet margin threshold used to forcefully widen the inter class distance is set to $\alpha = 0.3$. The hyperparameter controlling the strength of the information entropy penalty is set to $\lambda = 0.1$. The Softmax temperature is fixed at $\tau = 1.0$ for both training and inference, thereby avoiding any train–inference temperature mismatch. In the inference and testing phase, to maximize evaluation efficiency, only the final fusion feature $F_{final} \in \mathbb{R}^{1 \times C}$ output in Section 3.4 is extracted as the identity descriptor. A descending retrieval list is generated by calculating the Cosine Distance between feature vectors of the query set and the gallery set, and then the mAP and Rank-1 indicators are calculated.

(4) Constrained Evaluation Protocol for Real-Time Edge Deployment. To simulate resource-constrained edge computing environments and ensure a fair comparison, several standard performance-boosting techniques are strictly excluded from our evaluation protocols. Specifically, Side Information Embedding (SIE) for camera IDs and viewpoints is disabled during training to prevent the model from overfitting to specific camera topologies in open-world scenarios. Additionally, the Jigsaw Patch Module (JPM) and overlapping patch partition strategies are completely omitted to avoid high inference latency on edge hardware. For the testing phase, no K-reciprocal re-ranking postprocessing is applied, prioritizing raw system throughput (Frames Per Second, FPS). Under these identical, highly challenging real-time constraints, all baseline models and the proposed SPDR-Net are evaluated and compared fairly.

4.3. Experimental Results and Analysis

(i) Baseline Performance Evaluation in Standard Scenes. We first evaluate the fundamental feature representation capability of SPDR-Net on the Market-1501 and MSMT17 datasets. As summarized in Table 1, our method is compared with representative state-of-the-art baselines. Notably, the VLM-based CLIP-ReID achieves a slightly higher mAP of 83.1% and a Rank-1 accuracy of 93.8% on the clean Market-1501 single-scene benchmark. This marginal lead is reasonable, as CLIP-ReID leverages massive multimodal text–image pretraining datasets (e.g., LAION-400M) to align features, whereas SPDR-Net uses only visual supervision and an ImageNet-1K-pretrained backbone without external vision–language pretraining.

Table 1. Performance comparison on standard ReID datasets. The proposed method and the strongest comparison results discussed in the text are shown in bold.

Method	Model	Market1501 (Single) mAP/R-1	Msmt17 (Single) mAP/R-1	Market1501 (Multi) mAP/R-1	Msmt17 (Multi) mAP/R-1
CNN-based	PCB [11]	78.6/91.5	33.5/56.1	76.3/91.1	36.4/60.2
	ETNDNET(R50) [45]	80.1/92.0	35.0/57.0	77.5/91.5	37.9/61.4
Part-based	ReIDCaps [42]	81.8/92.8	36.5/58.5	79.2/92.0	38.8/62.5
	ViT-B [46]	79.5/91.8	34.5/56.5	77.0/91.3	37.0/60.5
ViT-based	TransReID [15]	81.0/92.5	35.8/57.8	78.5/91.8	38.2/61.8
	PASS [47]	81.3/92.1	36.2/58.2	79.2/91.5	39.0/62.4
	SOLIDER [48]	82.2/93.0	38.5/61.2	80.5/92.2	41.2/65.0
	PersonViT [49]	82.5/93.2	39.1/62.4	80.9/92.1	40.5/64.2
VLM-based	Pose2ID [50]	82.3/93.0	38.8/62.0	80.6/91.8	40.0/63.8
	CLIP-ReID [9]	83.1/93.8	40.2/64.5	81.5/92.8	41.8/65.8
Prompt-based	VersReID [8]	82.0/92.9	37.1/59.9	79.6/92.2	39.5/62.9
	PAD [18]	81.2/92.0	46.9/72.0	-	-
Ours	SPDR-Net	82.7/93.4	41.4/66.0	81.0/92.6	42.3/66.4

However, on the much larger and more complex MSMT17 dataset, which features drastic cross-camera illumination and viewpoint shifts, SPDR-Net achieves a superior mAP of 41.4% and a Rank-1 accuracy of 66.0% under our constrained single-scene setting, outperforming CLIP-ReID by 1.2% in mAP. Compared with the strong part-aware pretraining baseline PASS (36.2% mAP) and the semantic pretraining network SOLIDER (38.5% mAP), our model exhibits a clear performance leap. The newly included PAD baseline further shows that prompt-anchored vision–language distillation can be highly effective on recent ReID benchmarks. This comparison highlights that recent prompt-based ReID models are becoming stronger, while our method focuses on dynamic local routing for spatially asymmetric interference without relying on external text distillation.

To intuitively understand the performance of the model, Figure 7 visually displays the retrieval results using random query samples. Using the target query of MSMT17 (Target Sample, ID77) as an example, this pedestrian is located in an environment characterized by strong backlighting and posture changes. The compared methods retrieved incorrect identities (red boxes) at Rank-1 or even higher positions. This is because traditional models equivalently regard the dark part noise caused by backlighting as identity features. In contrast, SPDR-Net successfully hits the correct target images (green boxes) consecutively from Rank-1 to Rank-6. This indicates that the PGDR module keenly captures environmental distortions such as backlighting. The network dynamically down-weights the routing weights of the damaged parts, which improves the feature extraction ability of the model.

(ii) Robustness Analysis against Spatial-Asymmetric Interferences. To directly verify the primary objective of this work—coping with local asymmetric interferences—we

perform extensive comparative experiments on Occ-Duke (severe local occlusions) and SYSU-mm01 (visible–infrared cross-modality). The quantitative results are reported in Table 2. Under the Occ-Duke single-scene benchmark, SPDR-Net achieves the top-tier performance with 62.1% mAP and 75.6% Rank-1 accuracy, surpassing the powerful CLIP-ReID (61.2% mAP) and the recent PersonViT (60.5% mAP).

Table 2. Performance comparison on occluded and cross-modality datasets.

Method	Model	Occ-Duke (Single)	SYSU-mm01 (Single)	Occ-Duke (Multi)	SYSU-mm01 (Multi)
		mAP/R-1	mAP/R-1	mAP/R-1	mAP/R-1
CNN-based	PCB [11]	57.0/69.9	34.3/36.1	54.0/66.7	36.5/37.1
	ETNDNET(R50) [45]	59.4/72.9	36.3/38.0	57.0/69.7	38.5/39.8
Part-based	ReIDCaps [42]	60.4/73.9	37.3/39.0	58.0/70.8	39.5/40.8
	ViT-B [46]	57.9/71.4	35.0/36.9	56.5/69.2	38.0/39.3
ViT-based	TransReID [15]	60.0/73.5	36.9/38.6	57.2/70.0	38.6/39.8
	PASS [47]	58.1/71.4	35.8/37.4	55.4/68.2	37.2/38.5
	SOLIDER [48]	59.4/72.8	36.9/38.2	57.2/70.1	39.5/40.8
	PersonViT [49]	60.5/74.0	37.4/38.9	57.8/70.9	39.8/41.2
	Pose2ID [50]	60.3/73.8	37.2/38.6	57.5/70.6	39.6/41.0
VLM-based	CLIP-ReID [9]	61.2/74.8	38.5/39.8	58.9/71.8	41.2/42.9
Prompt-based	VersReID [8]	60.9/74.4	37.8/39.5	58.6/71.4	39.9/41.2
Ours	SPDR-Net	62.1/75.6	39.4/40.7	59.5/72.0	41.9/43.6

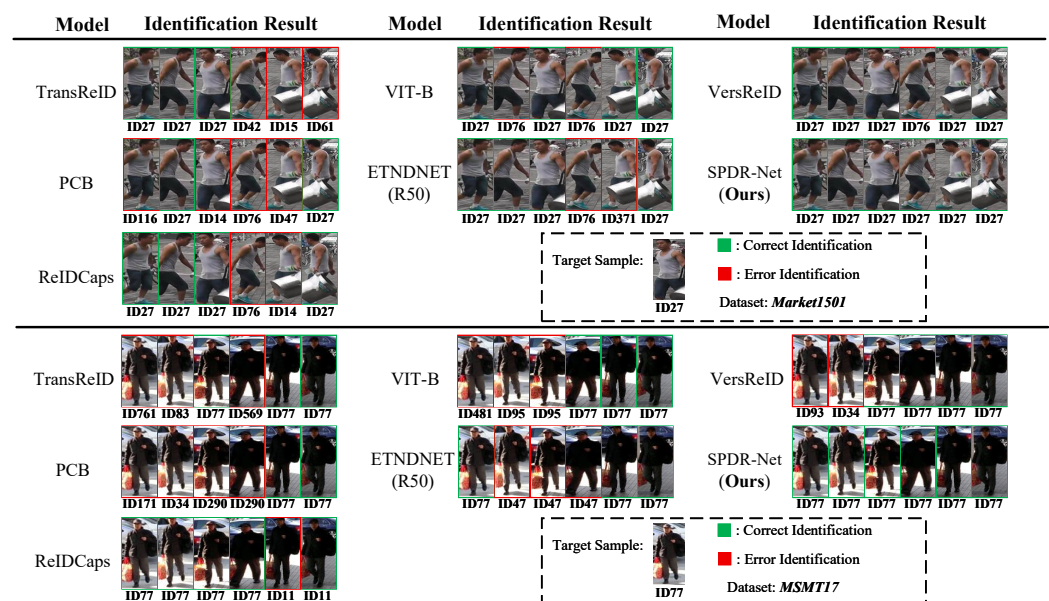


Figure 7. Visualization of performance test results on general ReID datasets.

This performance edge is highly convincing and highlights the theoretical necessity of our dynamic routing over dense global attention. In severe occlusion scenarios, models relying on global attention (such as TransReID, PASS, or CLIP-ReID) suffer from irreversible feature corruption because their global receptive fields inevitably mix the background noise of the occluders (e.g., vehicles, trees) into the identity descriptor. In contrast, our SPDR-Net utilizes the unsupervised body prior to segment the sequence into independent experts. Guided by the sparse entropy constraint, the PGDR network dynamically assigns an asymptotically negligible weight to the occluded expert path. This operation substantially suppresses noise propagation without losing the fine-grained details of the clear body parts. A similar advantage is observed on the cross-modality SYSU-mm01 dataset. Since infrared imaging strips away fine-grained clothing colors, models highly dependent on appearance

texture fail. Our PGDR network adaptively perceives this modal transition, prompting the GL-Fusion module to rely more on macroscopic structural contours rather than appearance details, thereby achieving robust and well-balanced cross-modality matching performance.

Figure 8 visualizes this result. In the occlusion test in the lower half (Target ID18), when the person's lower limbs are largely occluded, the baseline model mixes the stray texture of the occluder into the identity vector, resulting in a severe sorting error. However, SPDR-Net achieves accurate matching relying only on the unoccluded head and torso by substantially attenuating the feature weights of the damaged experts. In the infrared test in the upper half (Target ID54), SPDR-Net still maintains extremely high retrieval accuracy relying on its powerful global–local collaborative structural features, even when microscopic color features are lost.

(iii) Performance on Clothing Change Datasets. In long-term video surveillance spanning cross-periods, persons changing clothes poses a huge challenge for traditional ReID algorithms. Clothing changes completely alter the global color distribution of persons, causing models highly dependent on clothing color to fail completely. To test the model's ability to strip away color representations and extract essential identity features, we conduct strict evaluations on Celeb-ReID and PRCC, two datasets specifically designed for clothing-changing ReID. The results are shown in Table 3.

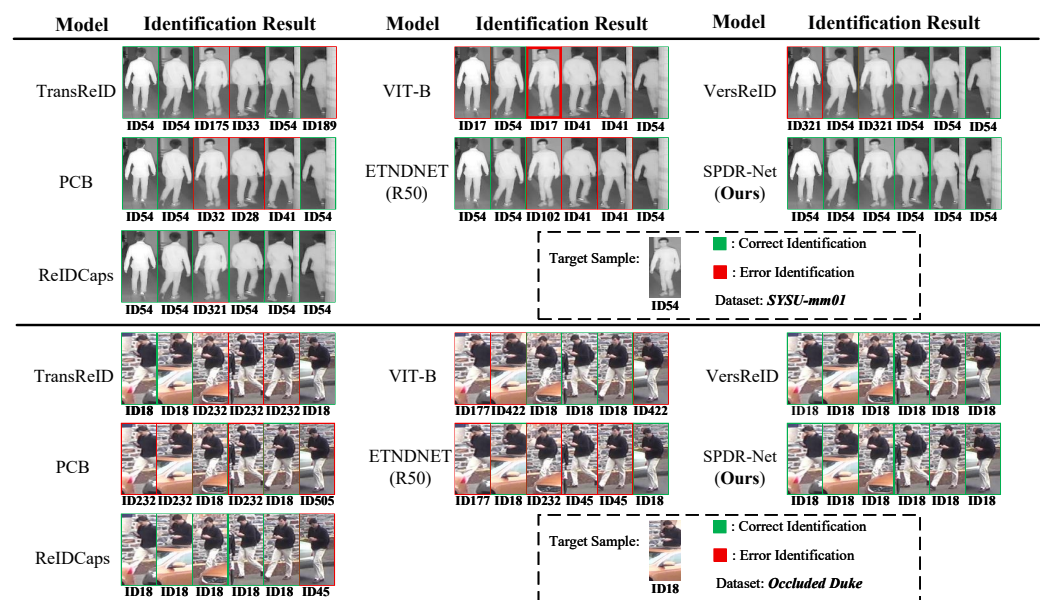


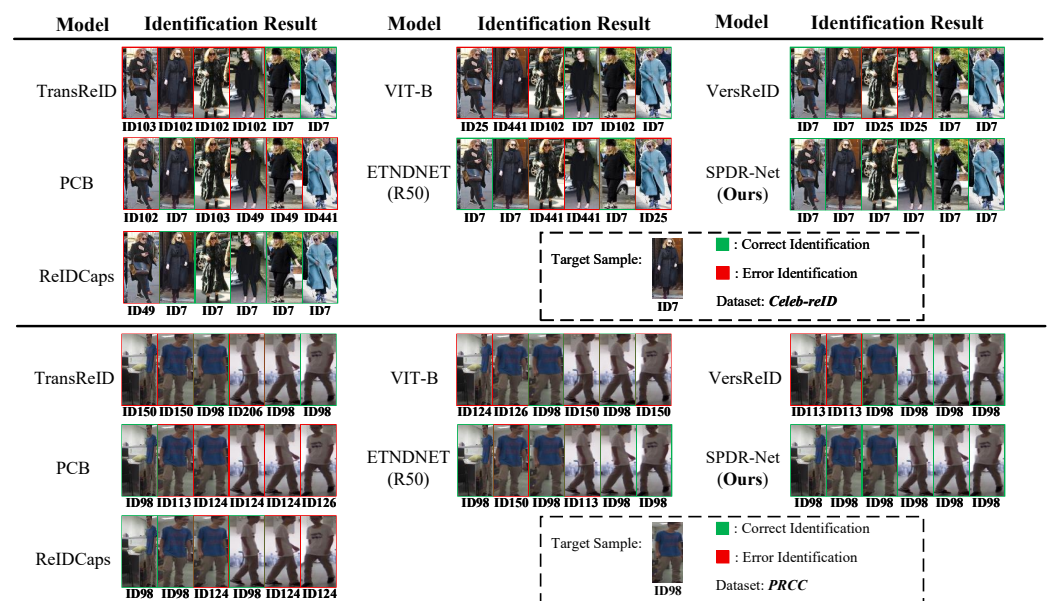
Figure 8. Visualization of retrieval performance under complex environmental constraints.

In the **single-scene** tests, SPDR-Net achieves 8.4% and 51.2% mAP, with 50.4% and 39.5% Rank-1 accuracy on Celeb-ReID and PRCC, respectively. Under **multi-scene** joint training, the performance further improves to 53.1% (PRCC mAP) and 49.4% (Celeb-ReID Rank-1). This robustness stems from the dynamic decoupling of global and local features. Even when clothing changes disrupt the global color anchor, local experts can still capture unchanged fine-grained attributes (e.g., facial contours or gait). Consequently, the PGDR and GL-Fusion modules synergistically shift the retrieval focus to these unaffected paths, thereby avoiding the dominance of color noise.

Table 3. Performance evaluation on clothing-changing person ReID datasets.

Method	Model	Celeb-ReID (Single)	PRCC (Single)	Celeb-ReID (Multi)	PRCC (Multi)
		mAP/R-1	mAP/R-1	mAP/R-1	mAP/R-1
CNN-based	PCB [11]	5.4/42.9	43.4/31.7	6.3/46.0	48.0/36.0
	ETNDNET(R50) [45]	6.5/45.3	46.0/34.5	6.9/47.0	49.0/37.0
Part-based	ReIDCaps [42]	7.1/46.3	47.0/35.5	7.5/48.0	50.0/38.0
	ViT-B [46]	6.2/44.8	45.5/34.0	6.7/46.6	48.7/36.7
ViT-based	TransReID [15]	6.9/46.0	46.8/35.2	7.0/47.1	48.5/36.7
	PASS [47]	6.5/45.2	45.2/33.1	6.8/45.9	47.5/35.8
	SOLIDER [48]	7.2/46.5	48.1/36.8	7.5/47.2	50.2/38.1
	PersonViT [49]	7.5/47.1	48.9/37.4	7.7/47.9	51.5/38.9
VLM-based	Pose2ID [50]	7.4/46.8	48.6/37.1	7.6/47.6	51.1/38.6
	CLIP-ReID [9]	7.8/48.2	49.5/38.2	8.1/48.9	52.4/39.9
Prompt-based	VersReID [8]	7.4/46.9	47.4/36.0	7.8/48.6	50.4/38.3
Ours	SPDR-Net	8.4/50.4	51.2/39.5	8.3/49.4	53.1/40.3

This dynamic compensation mechanism is visually demonstrated in Figure 9. Taking the target (Target ID7) of Celeb-ReID and the target (Target ID98) of PRCC as examples, it is clear that conventional baseline models are highly susceptible to the visual ambiguity of retrieving incorrect people wearing clothes of the same style and color. In contrast, SPDR-Net is not confused by global color mutations. It can cross clothing differences and stably lock onto the correct target with the same body topology and facial contour, showing a deep level of semantic understanding.

**Figure 9.** Qualitative evaluation on clothing-change ReID scenes.

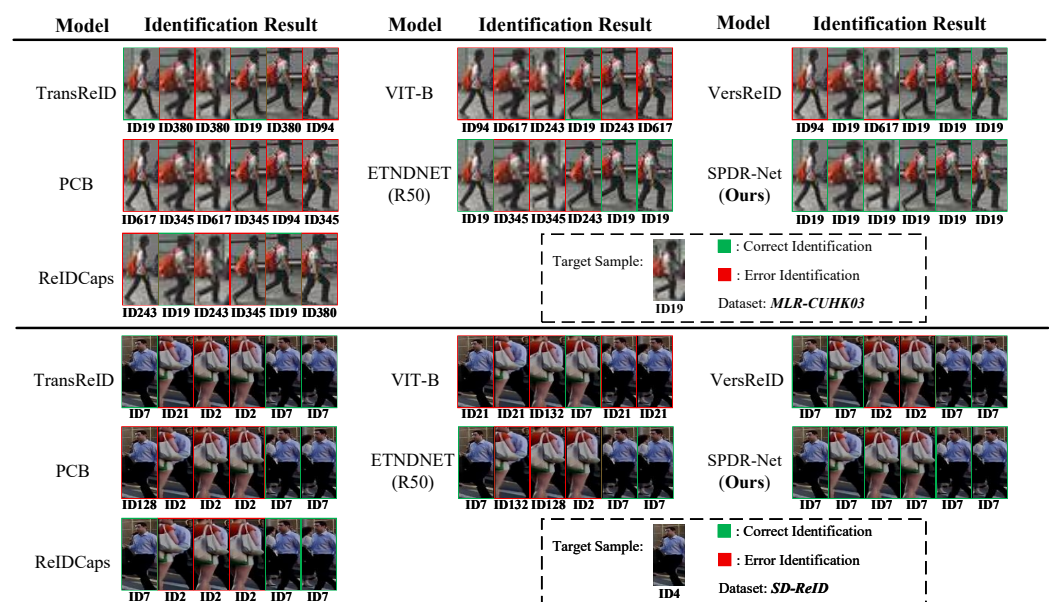
(iv) Performance on Low-Resolution and Construction Site Datasets. Public datasets are often isolated tests on a single interference variable. However, in real industrial deployment environments (such as complex city blocks and intersections), targets frequently face low resolution (loss of high-frequency texture caused by distant cameras). Corresponding evaluations are conducted on the low-resolution dataset (MLR-CUHK03) and our self-built block dataset (SD-ReID). The results are shown in Table 4.

Table 4. Performance evaluation on low-resolution and street ReID datasets.

Method	Model	MLR-CUHK03 (Single)	SD-ReID (Single)	MLR-CUHK03 (Multi)	SD-ReID (Multi)
		mAP/R-1	mAP/R-1	mAP/R-1	mAP/R-1
CNN-based	PCB [11]	80.3/85.2	92.1/96.0	79.8/84.7	94.6/97.2
	ETNDNET(R50) [45]	84.6/89.4	95.2/98.2	83.9/88.1	97.1/99.0
Part-based	ReIDCaps [42]	86.2/91.1	96.1/98.8	85.7/90.0	98.6/99.8
	ViT-B [46]	82.8/87.5	94.0/97.3	81.4/86.3	95.9/98.1
ViT-based	TransReID [15]	85.5/90.3	95.7/98.5	84.8/89.2	98.0/99.5
	PASS [47]	83.1/87.2	93.5/96.8	82.1/86.3	95.8/97.9
	SOLIDER [48]	84.8/89.1	94.8/97.5	84.2/88.5	97.4/98.9
	PersonViT [49]	85.9/90.4	95.5/98.0	85.1/89.4	98.0/99.1
	Pose2ID [50]	85.6/90.1	95.2/97.8	84.8/89.1	97.8/98.9
VLM-based	CLIP-ReID [9]	86.5/91.2	96.0/98.4	86.0/90.2	98.5/99.4
Prompt-based	VersReID [8]	85.7/90.9	95.8/98.8	84.9/89.3	98.3/99.8
Ours	SPDR-Net	87.7/91.9	96.5/98.9	87.1/91.0	99.0/99.9

On the MLR-CUHK03 dataset, SPDR-Net achieves a performance of 87.7% (mAP) and 91.9% (Rank-1). On the SD-ReID dataset, it achieves 96.5% mAP and 98.9% Rank-1 (single-scene). Furthermore, under **multi-scene** conditions, it reaches an excellent performance of 99.0% (mAP) and 99.9% (Rank-1). When a certain part suffers from pixel level blur due to low resolution, or physical occlusion caused by non-motorized vehicles in the block, the PGDR module prompts the model to adaptively retain and strengthen the purest local paths. Consequently, the model extracts pure identity representations under a high-noise background.

The visualization results in Figure 10 illustrate this. In the upper half of the MLR-CUHK03 dataset test (Target ID19), although the image is highly blurred, the SPDR-Net framework still accurately identifies the target relying on the macroscopic topological structure. In the lower half of the SD-ReID dataset test (Target ID4), facing intersecting persons, the SPDR-Net framework robustly mitigates occlusion interference and achieves accurate recognition.

**Figure 10.** Retrieval results on low-resolution benchmarks.

In practical closed-loop video surveillance applications, the system not only relies on the accuracy of the Rank-1 hit but also requires the top 5 or top 10 candidate results to contain a high accuracy rate to support secondary matching or manual review. For this

reason, Figure 11 further compares the Rank-5 (R-5) and Rank-10 (R-10) retrieval metrics of the models across all scenes.

The results show that SPDR-Net achieves superior performance on both deep retrieval metrics. This indicates that through global and local feature residual fusion, the identity features extracted by SPDR-Net have high intra-class compactness in the metric space. This compactness tightly clusters the candidate images of the correct target. The high accuracy of this deep candidate list verifies the adaptability of the model to the environment under massive open data.

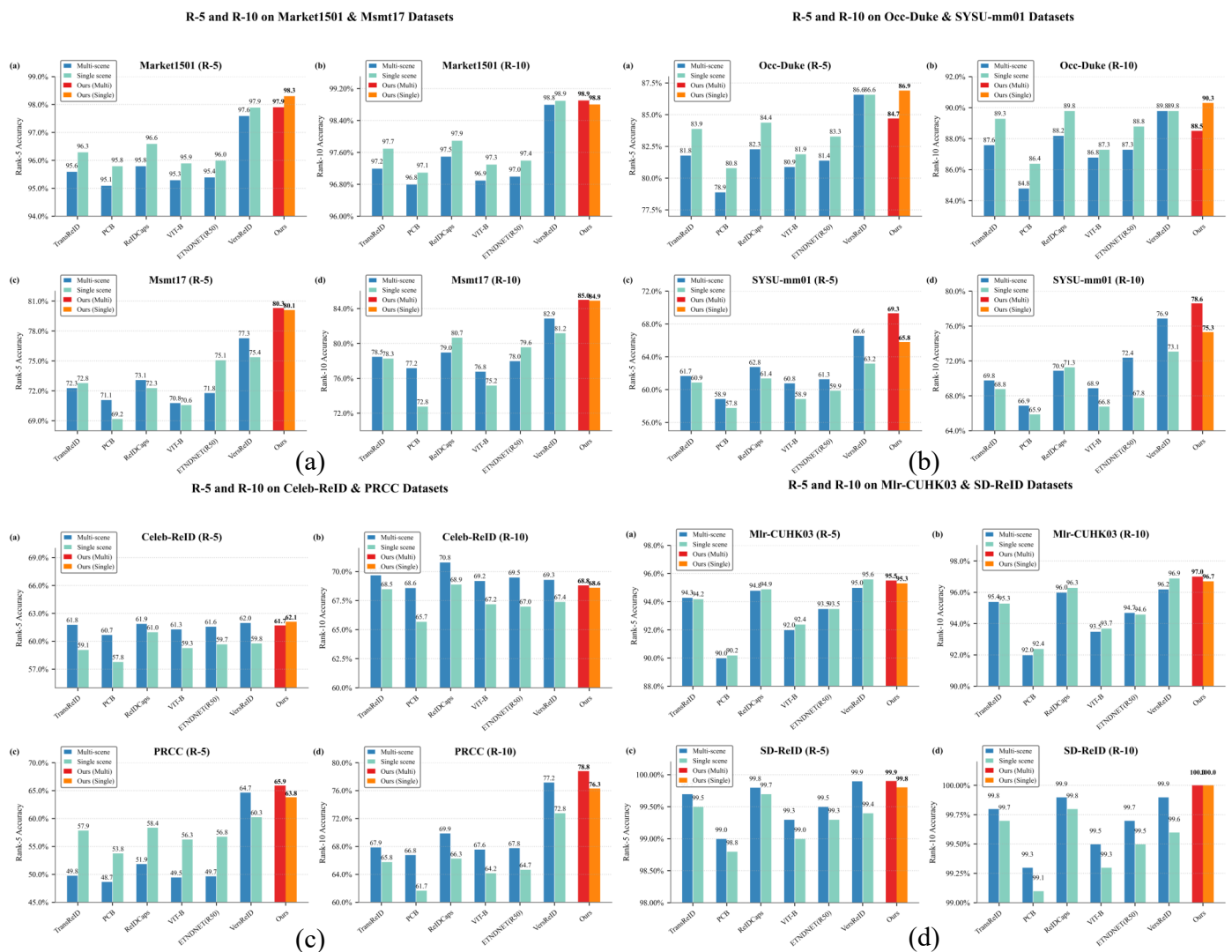


Figure 11. Comparison of Rank-5 and Rank-10 retrieval metrics between SPDR-Net and baselines: (a) Market1501 and MSMT17; (b) Occ-Duke and SYSU-mm01; (c) Celeb-ReID and PRCC; and (d) MLR-CUHK03 and SD-ReID. Within each group, the four small panels show Rank-5 and Rank-10 results for the two named datasets, in left-to-right and top-to-bottom order.

4.4. Ablation Study

To deeply verify the effectiveness of each core module in SPDR-Net and its contribution to the overall performance, rigorous ablation experiments were conducted. Specifically, the SPDR-Net network is decoupled into three main parts: the semantic part expert construction and refinement module (denoted as SPCR), the scene-prompt-guided dynamic routing module (denoted as PGDR, covering the PGDR mechanism), and the global-local feature collaborative fusion module (denoted as GL-Fusion). This ablation experiment conducted detailed combination tests on eight benchmarks, including normal, occlusion,

cross-modality, and our self-built datasets. All tests adopted the multi-scene joint training paradigm to impose the maximum domain shift pressure. The results are shown in Table 5. To clarify the ablation baseline, when all three proposed modules (SPCR, PGDR, and GL-Fusion) are disabled, the network structure simplifies exactly to the baseline VersReID model. Consequently, the performance in the first row of Table 5 matches the results of VersReID in Tables 1–4. To make this clear, we have labeled this row as “Baseline (VersReID)” in Table 5.

Table 5. Performance evaluation of ablation experiments (Unit: %).

SPCR	PGDR	GL-Fusion	Market1501		Msmt17		Occ-Duke		SYSU-mm01	
			mAP	R1	mAP	R1	mAP	R1	mAP	R1
✓	Baseline (VersReID) ¹		79.6%	92.2%	39.5%	62.9%	58.6%	71.4%	39.9%	41.2%
	✓		77.2%	90.1%	37.8%	60.5%	56.5%	69.8%	38.5%	40.1%
			78.5%	91.0%	38.9%	61.8%	57.8%	70.5%	39.2%	40.8%
✓	✓	✓	80.1%	92.3%	37.2%	60.1%	55.9%	69.2%	38.1%	39.9%
		✓	81.0%	92.6%	42.3%	66.4%	59.5%	72.0%	41.9%	43.6%
	✓	✓								
SPCR	PGDR	GL-Fusion	Celeb-ReID		PRCC		Mlr-CUHK03		SD-ReID	
			mAP	R1	mAP	R1	mAP	R1	mAP	R1
✓	Baseline (VersReID) ¹		7.8%	48.6%	50.4%	38.3%	84.9%	89.3%	98.3%	99.8%
	✓		7.4%	47.0%	49.8%	37.5%	85.2%	89.8%	98.0%	99.7%
			7.6%	47.8%	50.5%	38.0%	85.5%	90.2%	98.7%	99.8%
✓	✓	✓	7.3%	46.8%	49.2%	37.2%	84.8%	89.5%	98.3%	99.6%
		✓	8.3%	49.4%	53.1%	40.3%	87.1%	91.0%	99.0%	99.9%
	✓	✓								

¹ The performance metrics reported in the rows “Baseline (VersReID)” are cited baseline scores from [8] for structural reference, rather than results newly computed through our ablation experiments.

From the data results in Table 5, it can be observed that when a certain module is intervened individually, the model performance does not significantly surpass the baseline. The fundamental reason is that the three-stage architecture of SPDR-Net presents highly coupled and complementary characteristics. The SPCR module provides the necessary local spatial division coordinate system and basic feature denoising for the model. The PGDR module relies on the spatial division of the SPCR module to execute core weight allocation by perceiving the environmental context. Meanwhile, the GL-Fusion module provides macroscopic identity constraints for local features to prevent them from falling into the trap of over-localization. Missing any of these links will cause feature representation degradation and even introduce additional structural noise. Only when the three modules work synergistically can the network stimulate a complete “structural prior-part refinement-global complementarity” effect, allowing all indicators to reach their optimum.

To further explain the actual mechanism of scene-prompt-guided dynamic gating routing under composite interference from a visual perspective, this paper utilizes Grad-CAM technology. We perform visual mapping on the feature attention weights of each part expert in the network fusion layer. This operation generates the attention tendency heat maps, as shown in Figure 12.

From the distribution of the red highlighted areas in Figure 12, it can be clearly observed that SPDR-Net can achieve precise focusing on valid identity features. For example, in the occlusion samples of the Occ-Duke dataset, when the lower half of the person is physically occluded over a large area, the model’s attention automatically avoids the occluder and highly concentrates on the undamaged upper body and head. In the infrared cross-modality scene of SYSU-mm01, the attention adaptively shifts from the failure-prone clothing color to the overall physical contour of the person. This visualization result proves

that the network possesses the decision-making ability to dynamically judge environmental interference and strongly suppress affected feature paths.

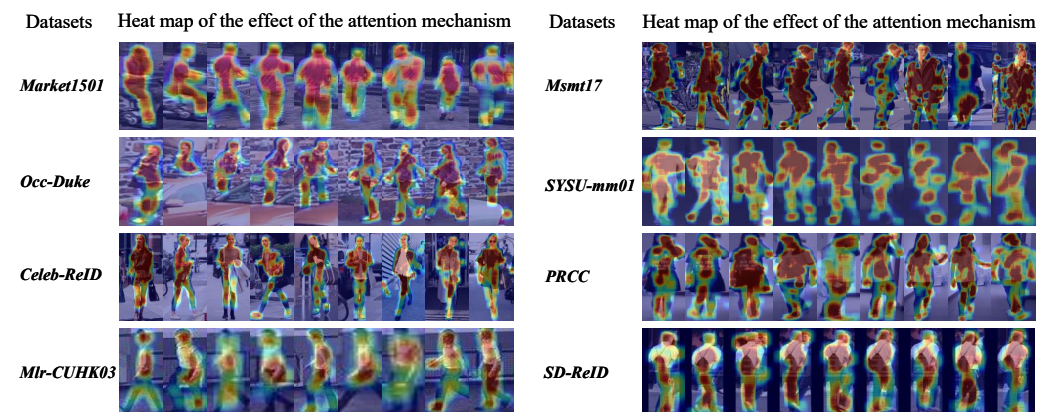


Figure 12. Visualization of expert feature attention tendencies of SPDR-Net when facing various spatially asymmetric interferences. Warm colors (red and yellow) indicate high attention responses, whereas cool colors (green and blue) indicate lower responses.

In addition to validating test-time generalization metrics (mAP and Rank-1), evaluating the training convergence behavior across diverse environmental attributes is essential to verify the optimization stability of the proposed SPDR-Net. To this end, we record and compare the joint loss curves of SPDR-Net and recent state-of-the-art ReID methods (including PASS [47], SOLIDER [48], PersonViT [49], and CLIP-ReID [9]) over training epochs on all eight benchmarks in Figures 13 and 14. We explicitly clarify that a lower training loss magnitude does not always imply superior test-time retrieval generalization. However, the convergence rate and curve smoothness strongly reflect the gradient properties and optimization landscape. As illustrated in the curves, the multimodal CLIP-ReID baseline exhibits the lowest initial and final loss values across all eight datasets. This fast convergence is mathematically reasonable. CLIP-ReID leverages massive pre-aligned text–image knowledge from foundation models, giving it an extremely strong starting point in the latent space.

Among the pure visual models trained from scratch, the proposed SPDR-Net consistently exhibits a faster descent rate and smaller oscillation amplitudes compared to traditional baselines (e.g., PCB [11], TransReID [15], and SOLIDER [48]). This excellent optimization stability remains highly consistent under different domain-shift challenges. On the massive MSMT17 dataset, which features severe cross-camera illumination variations, and the MLR-CUHK03 dataset, which is characterized by heavy resolution degradation, SPDR-Net rapidly converges to a stable state within the first 25 epochs. This superior optimization behavior is mathematically attributed to two core designs: (1) The learnable scene prompt prior provides a highly generalized coordinate initialization, suppressing early gradient volatility during domain transitions. (2) The PGDR module adaptively down-weights the feature activations of heavily corrupted parts. In the backpropagation phase, this operation strongly attenuates the gradients propagated through highly conflicting or noisy expert paths. By shielding the clean experts from the dominant influence of noisy gradients, our routing mechanism effectively smooths the joint optimization trajectory across diverse datasets, steering the network toward a more stable and robust local minimum without requiring the astronomical parameter overhead of foundation models.

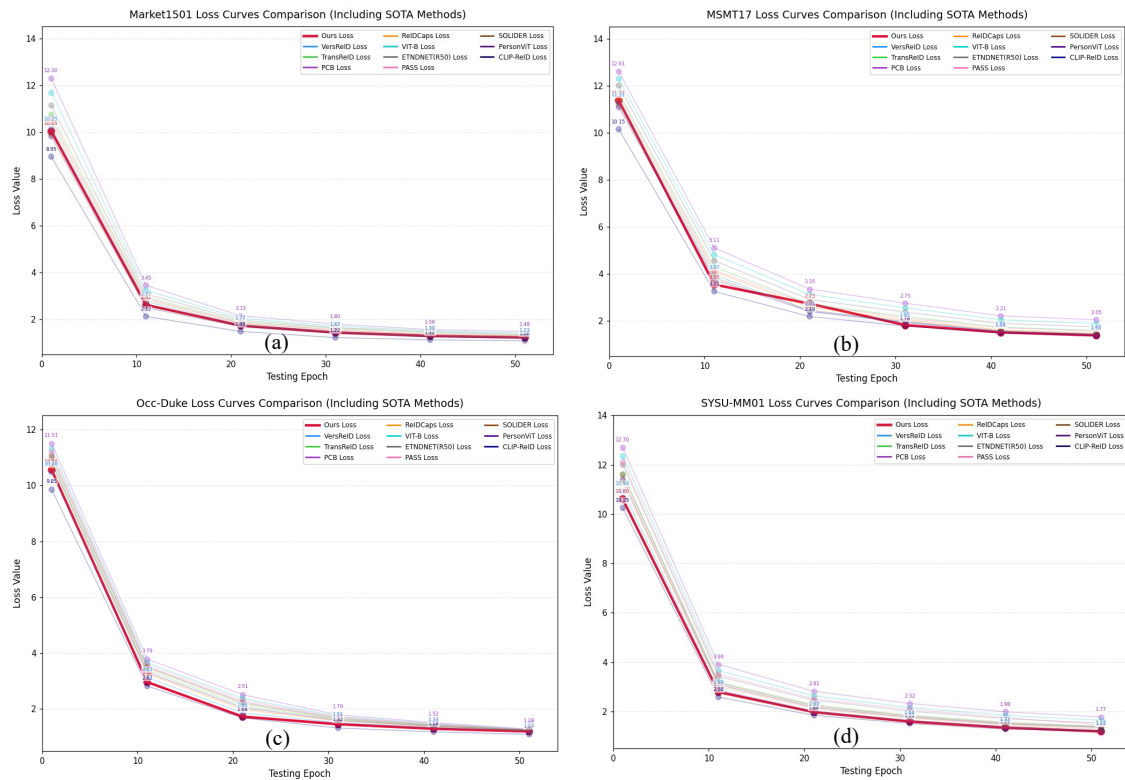


Figure 13. Comparison of training loss convergence curves between SPDR-Net and baseline models across various benchmark datasets: (a) Market1501; (b) MSMT17; (c) Occ-Duke; and (d) SYSU-mm01. Line colors identify the compared models as listed in each panel legend.

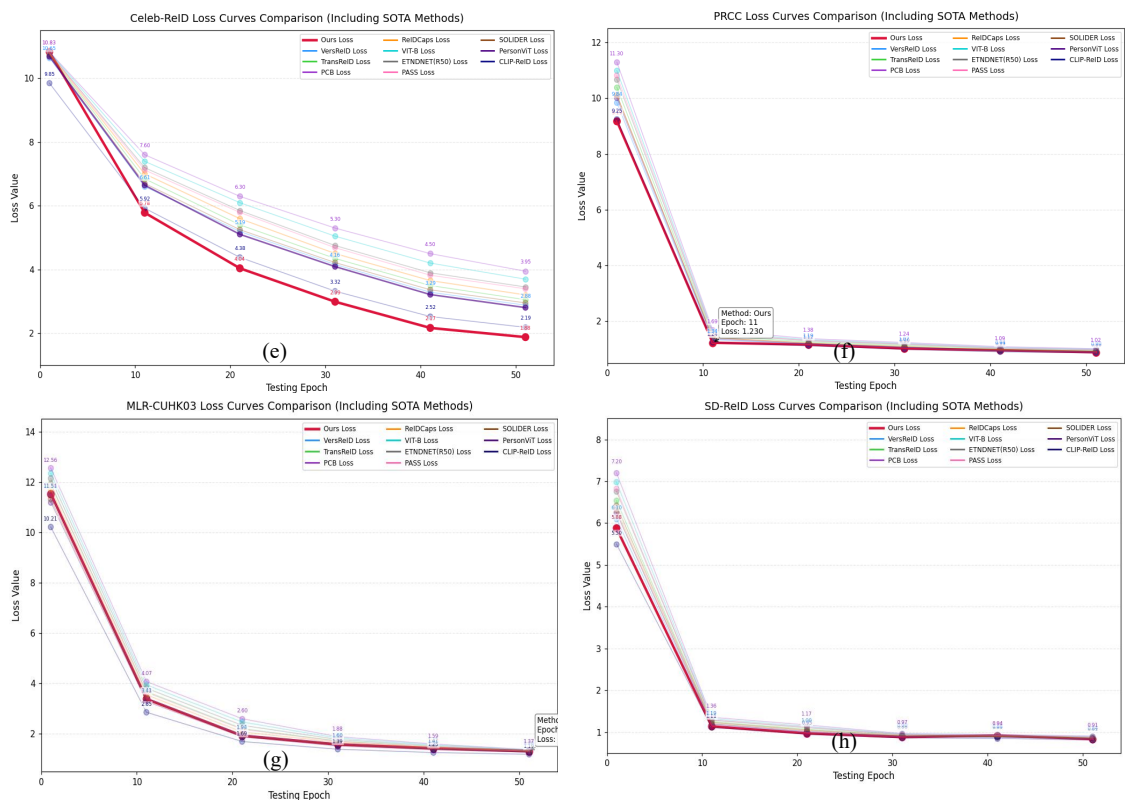


Figure 14. Comparison of training loss convergence curves between SPDR-Net and baseline models across various benchmark datasets: (e) Celeb-ReID; (f) PRCC; (g) MLR-CUHK03; and (h) SD-ReID. Line colors identify the compared models as listed in each panel legend.

4.5. Fine-Grained Evaluation on the SD-ReID Benchmark

To fully explore the benchmark evaluation value of the SD-ReID dataset and deeply analyze the real performance of the proposed model when dealing with different types of spatially asymmetric interference, this section conducts a deeper breakdown evaluation of the SD-ReID dataset.

First, based on the open-world challenge characteristics elaborated previously, the query samples in the SD-ReID test set are divided in the SD-ReID test set (containing 2055 images in total) into four specific interference attribute subsets. **Normal**: Contains person samples under standard lighting, without obvious occlusion, and with normal resolution. **Occlusion**: The torso or lower limbs of the target person are asymmetrically occluded by non-motorized vehicles, street signs, or other persons in the block. **Low-Resolution**: Due to the large depth of field, the target pixels are blurred, lacking microscopic textures. **Illumination**: The target is in environments with strong indoor/outdoor backlighting, local tree shadows, or dim nighttime lighting. Under the multi-scene joint training paradigm, we conduct a fine-grained mAP performance comparison between SPDR-Net and the currently best performing baseline model (VersReID). The specific test results are shown in Table 6.

Table 6. Fine-grained performance comparison (mAP) under different environmental attribute subsets of the SD-ReID dataset.

Scene Subset	Image Count	Baseline	SPDR-Net (Ours)	Improvement
Normal	855	99.1%	99.6%	+0.5%
Low-Resolution	400	97.2%	98.7%	+1.5%
Illumination	300	96.5%	98.3%	+1.8%
Occlusion	500	95.4%	98.5%	+3.1%
Total Set	2055	98.3%	99.0%	+0.7%

In the “Normal” subset, the baseline model reaches a performance saturation of 99.1% due to minimal environmental noise. However, SPDR-Net still achieves a 0.5% improvement. This gain is attributed to the SPCR module, which effectively filters background clutter even without obvious interference. Secondly, in the “Low-Resolution” and “Illumination” subsets, the baseline performance degrades by about 2% due to the loss of microscopic textures and color consistency. Conversely, SPDR-Net yields improvements of 1.5% and 1.8%, respectively. This benefits from the GL-Fusion module. When local details are blurred, GL-Fusion leverages global macroscopic contours for residual compensation to resist domain shifts. Finally, in the “Occlusion” subset, the baseline mAP drops to 95.4% because global prompts fail to isolate occluder noise. In contrast, SPDR-Net achieves a 3.1% improvement. This significant gain is achieved because the PGDR module adaptively estimates local asymmetric interference and down-weights the occluded expert paths. Consequently, it mitigates the influence of localized noise and emphasizes reliable features from unoccluded areas.

The previous tests were all conducted under a domain-consistent setting. However, the biggest challenge faced by open-world ReID is unpredictable out-of-distribution scenes. In practical applications, the system often needs to be deployed directly without obtaining the training data of the target site. To verify that SPDR-Net possesses a true cross-domain anti-interference mechanism rather than just overfitting to specific backgrounds, an extremely challenging cross-domain zero-shot generalization experiment is designed. In this experiment, we select the currently largest public ReID dataset, MSMT17, as the source domain for model training. Then, without any fine-tuning, we directly use the model to perform inference and retrieval on the completely unknown SD-ReID testing set (target

domain). A horizontal comparison is conducted among mainstream models representing four different technical paradigms. The evaluation results are shown in Table 7.

Table 7. Cross-domain zero-shot generalization performance comparison on the SD-ReID dataset benchmark.

Technical Paradigm	Model	mAP (%)	Rank-1 (%)
Static local stitching	PCB [11]	32.4%	45.1%
Static global receptive field	ViT-B [46]	41.7%	53.6%
Global prompt learning	VersReID [8]	48.2%	61.5%
Local dynamic routing (Ours)	SPDR-Net (Ours)	56.8%	70.2%

Table 7 shows a severe performance drop for all models during direct cross-domain inference, highlighting the unique environmental complexity of the SD-ReID dataset. Static models perform poorly (mAPs of 32.4% and 41.7%). They overfit to the source domain (MSMT17) background and use fixed weights, which mistakenly forces unknown SD-ReID noises into the feature space. Conversely, VersReID increases the mAP to 48.2%, proving that prompt learning aids macroscopic domain alignment. Ultimately, SPDR-Net achieves the best performance with 56.8% mAP and 70.2% Rank-1. Instead of relying on background memory, the PGDR module utilizes the sparse entropy constraint (L_{ent}) to estimate expert reliability and down-weight less reliable local expert paths in real time. This dynamic adaptive mechanism endows SPDR-Net with excellent cross-domain anti-interference capabilities.

To address the potential concern regarding the high performance on the SD-ReID dataset, we provide a detailed analysis of its protocol. The SD-ReID dataset contains 147 identities, which is relatively small compared to MSMT17. This smaller scale naturally reduces retrieval confusion in closed-loop matching. Additionally, all images are automatically cropped by a high-precision YOLO detector, ensuring excellent spatial alignment. To perform a more rigorous evaluation, we introduce a strict Cross-Camera and Cross-Time (CCCT) testing protocol. Under this protocol, the query set and the gallery set are divided such that the query images of a target person and their corresponding gallery images are captured not only by different camera views but also at completely different time periods. This temporal and spatial separation reduces the threat of short-term frame redundancy and provides a highly challenging benchmark for real-world deployment.

To evaluate the generalization limit of SPDR-Net under the newly introduced Cross-Camera and Cross-Time (CCCT) protocol, we conduct a comparative experiment on the SD-ReID dataset. As shown in Table 8, when subjected to this rigorous spatial and temporal separation, all models experience a visible performance drop compared to the standard setting. This drop successfully verifies the high difficulty of our CCCT protocol. Under this challenging setting, the proposed SPDR-Net still achieves 88.1% mAP and 92.4% Rank-1, outperforming the state-of-the-art VersReID by 3.5% in mAP. This result firmly demonstrates the robust generalization capability of our dynamic routing mechanism rather than simple background overfitting.

Table 8. Performance comparison under the Cross-Camera and Cross-Time (CCCT) protocol on the SD-ReID dataset.

Method	Model	mAP (%)	Rank-1 (%)	Rank-5 (%)
Part-based	PCB [11]	73.5	79.1	86.4
Transformer	ViT-B [46]	79.2	84.8	91.0
Prompt-based	VersReID [8]	84.6	89.2	94.5
Ours	SPDR-Net	88.1	92.4	96.2

4.6. Sensitivity Analysis of Core Hyperparameters

To deeply explore the internal state of the Prompt-Guided Dynamic Routing (PGDR) module during actual operation, and to verify the robustness of the model to the underlying topology and key hyperparameters, this section conducts a detailed sensitivity analysis on two core parameters: the number of local experts K , and the sparse routing entropy loss weight λ . To test the role of these two parameters in dealing with spatially asymmetric interference, the experiments in this section are conducted on the Occ-Duke dataset (containing severe partial occlusion) and the self-built SD-ReID dataset.

1. Topological granularity analysis of the number of local experts K . In the semantic part expert construction and refinement (SPCR) module, parameter K defines the number of independent semantic experts into which the person spatial topology is divided. K determines the receptive field area of the model for perceiving local interference. It also dominates the tradeoff between feature semantic fine granularity and spatial position robustness. To explore the global optimal solution of this tradeoff, K is set to values in $\{1, 3, 4, 6, 8, 12\}$. The corresponding performance trend is presented in Figure 15.

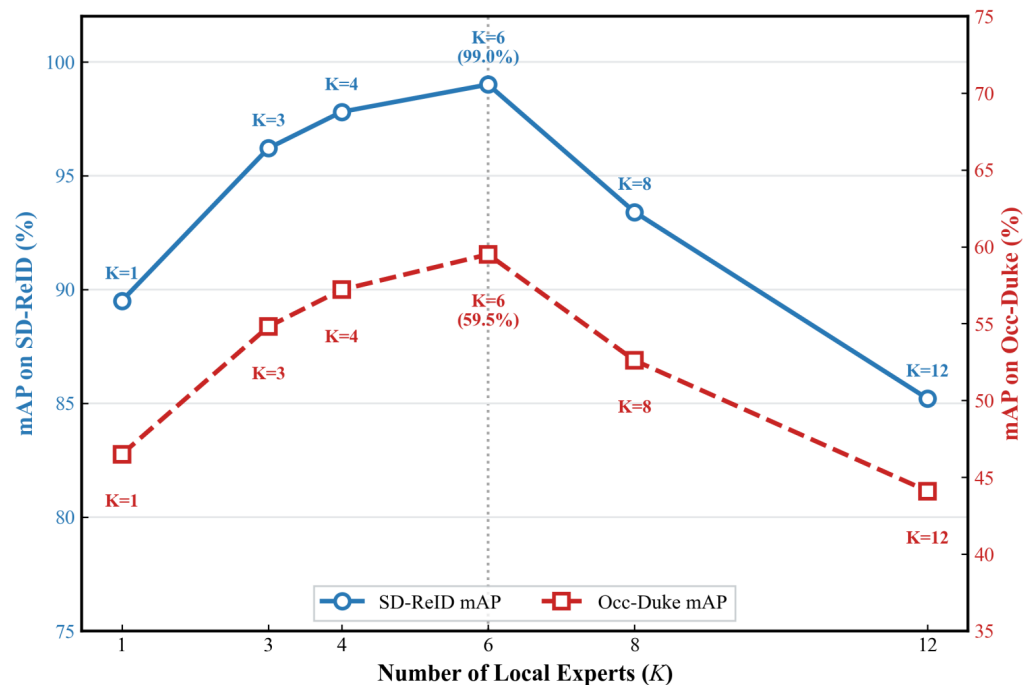


Figure 15. Sensitivity analysis of the topological granularity regarding the number of local experts K .

When $K = 1$, the model degenerates into pure global feature extraction and loses the ability to combat spatially asymmetric interference. This leads to the lowest mAP. With the initial increase of K (e.g., $K = 3$), the model performance recovers significantly. However, $K = 3$ is not the optimal solution. Due to the coarse segmentation granularity, when a tiny part of the person (e.g., only the calf) is occluded, the PGDR module has to down-weight the entire broad lower-body path (including the clean features of the thigh region) to avoid noise in that area. This inevitably leads to collateral damage to valid local semantic features. When K continues to increase and reaches $K = 6$, the model's performance reaches its peak. For an input image with a height of 384 pixels, $K = 6$ allows each local expert to be exactly responsible for a continuous receptive field with a height of about 64 pixels. This division granularity enables a good balance between “feature semantic fine-grainedness” and “spatial position robustness”. On the one hand, the sufficiently fine segmentation enables the gating network to strongly suppress localized polluted features. On the other hand, $K = 6$ ensures that each expert path contains enough context information internally.

This context supports the self-attention mechanism within the framework to perform effective Local Feature Refinement.

However, when the value of K expands excessively (e.g., $K = 8$ or even $K = 12$), all indicators begin to decline. Explained from a deep mechanism perspective, excessively detailed division triggers three problems. First, Semantic Fragmentation: Undersized image patches may only contain solid-colored fabrics of a single color, completely losing the information entropy required to distinguish different identities. Second, severe spatial misalignment: In the uncontrolled open environment, person posture changes and slight bounding box jitters will cause semantic mismatching of these tiny segmented parts between two images. Third, Routing Dilution: Too many part expert branches greatly smooth the probability distribution of the Softmax function. This weakens the intervention strength of the sparse entropy loss (L_{ent}). Consequently, the PGDR module loses its decision-making ability to strongly suppress polluted feature paths.

2. Decision strength analysis of the sparse routing entropy loss weight λ . In the joint optimization objective, the hyperparameter λ is responsible for balancing the relationship between the identity metric feature space and the routing decision certainty. λ determines whether the network has the ability to strongly suppress the feature paths of damaged parts when facing interference. We set $\lambda \in \{0, 0.01, 0.05, 0.1, 0.5, 1.0\}$ for testing. The performance trend is shown in Figure 16.

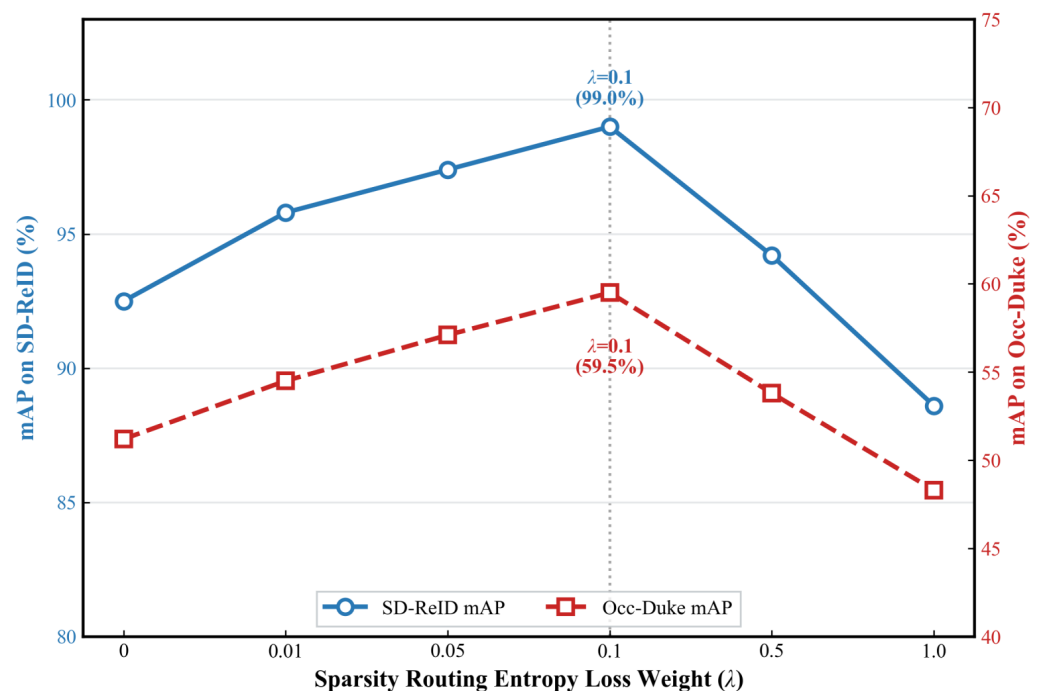


Figure 16. Sensitivity analysis of the decision strength regarding the sparsity routing entropy loss weight λ .

To provide quantitative evidence for the routing-weight distribution and to clarify the boundary of the entropy penalty, we further compute the mean Shannon entropy of the routing weights on the Occ-Duke testing set. Clean samples and occluded samples are evaluated separately under different values of λ . For the adopted $K = 6$ expert configuration, the maximum theoretical entropy is $\ln(6) \approx 1.792$. The statistical results are reported in Table 9.

As shown in Table 9, when $\lambda = 0$ (i.e., no entropy penalty), the routing distributions remain close to uniform for both clean and occluded samples, and the entropy gap is small ($\Delta H = 0.139$). This explains why the PGDR module tends to make ambiguous

routing decisions without explicit information-theoretic pressure. When $\lambda = 0.1$, the model achieves the best performance in Figure 16. At this setting, clean samples still maintain a high mean entropy ($H_{clean} = 1.692$), indicating that the router largely preserves balanced body-part information in non-occluded cases. In contrast, the entropy of occluded samples decreases to $H_{occ} = 1.284$, producing the largest clean–occluded gap ($\Delta H = 0.408$). This result supports the interpretation that selective polarization is an emergent outcome of the joint optimization rather than a hard noise-triggered switch. When $\lambda = 0.5$, the entropy of clean samples collapses sharply to 0.512, which empirically confirms the risk of blind polarization. Therefore, the degradation observed in Figure 16 at large λ is consistent with the reviewer’s concern: an overly strong entropy penalty can suppress useful normal expert features even in clean scenes.

Table 9. Mean routing entropy statistics on the Occ-Duke dataset.

Entropy Penalty (λ)	Clean Samples (H_{clean})	Occluded Samples (H_{occ})	Delta (ΔH)
$\lambda = 0.0$	1.754	1.615	0.139
$\lambda = 0.1$ (Optimal)	1.692	1.284	0.408
$\lambda = 0.5$ (Over-polarized)	0.512	0.315	0.197

4.7. Computational Complexity and Real-Time Inference Efficiency

To verify the practical utility of SPDR-Net, we conduct a detailed computational complexity analysis. The parameter size, FLOPs, single-frame inference latency, and throughput (Frames Per Second, FPS) of different models are summarized in Table 10. All models are evaluated under identical hardware configurations on a single NVIDIA RTX 4090D GPU. As reported in Table 10, lightweight CNN models like OSNet and part-aware models like PCB exhibit fast inference speeds (1818 FPS and 1204 FPS, respectively). However, their retrieval accuracies in complex scenes are unsatisfactory, as shown in Tables 1–4. On the other hand, state-of-the-art Transformer-based models (such as TransReID, VersReID, and CLIP-ReID) achieve higher accuracies but require massive global self-attention computations. This quadratic complexity pushes their FLOPs above 17G and reduces their inference speeds to under 400 FPS, which fails to meet real-time requirements.

Table 10. Comparison of computational complexity and real-time inference efficiency.

Method	Model	Params (M)	FLOPs (G)	Latency (ms/img)	Throughput (FPS)
CNN-based	OSNet [51]	2.2	1.0	0.55	1818
	PCB [11]	27.1	4.8	0.83	1204
	ViT-B [46]	85.8	16.8	2.37	421
ViT-based	TransReID [15]	86.4	17.1	2.53	395
	PASS [47]	86.8	16.9	2.45	408
	SOLIDER [48]	87.2	17.0	2.49	402
	PersonViT [49]	92.5	18.2	2.76	362
VLM-based	CLIP-ReID [9]	142.1	22.4	3.51	285
Prompt-based	VersReID [8]	88.1	17.5	2.63	380
Ours	SPDR-Net	87.3	14.1	1.81	552

In contrast, the proposed SPDR-Net achieves an outstanding balance between accuracy and computational speed. Although the parameter size of SPDR-Net (87.3M) is comparable to the baseline ViT-B, the total computational cost (FLOPs) is significantly compressed from 16.8G to 14.1G. This optimization is achieved because our part-aware ‘divide-and-conquer’ attention strategy (Section 3.2) partitions the long patch sequence into three independent, parallel streams. This parallel design reduces the attention computational complexity from $\mathcal{O}(N^2C)$ to $\mathcal{O}(N^2C/3)$. Consequently, the system throughput of SPDR-Net reaches

552 FPS (a 45.2% speedup over VersReID), with a low latency of only 1.81 ms per image. This evaluation confirms the outstanding potential of the proposed model for real-time edge deployment.

5. System Construction and Application

Although the previous sections have verified the retrieval capability of SPDR-Net on multiple static gallery datasets, there is still a gap between pure static image matching and real-world industrial deployment. In a real open-world video surveillance network, the system faces more complex dynamic challenges: 1. Person features are no longer perfectly cropped images, but highly depend on the dynamic bounding boxes generated by the front-end object detector. The jitter and truncation of the boxes introduce additional structural noise. 2. Continuous tracking across indoor and outdoor cameras is accompanied by intense and unpredictable illumination mutations and posture transitions. To conduct end-to-end verification of the proposed theoretical model toward practical ReID engineering systems, we built a multi-scene software and hardware testing system. This system integrates front-end detection and back-end ReID.

As shown in Figure 17, this test system consists of a 75-inch UHD (3840 × 2160) master control terminal and three topologically distributed cameras (including one indoor corridor bullet camera and two outdoor environment dome cameras). This realizes real-time collection, processing, and rendering of video streams. To strictly control variables and minimize the interference of the detector on the final indicators, we uniformly employ the state-of-the-art YOLO (specifically, YOLOv12x) as the front-end object detector. This ensures that all comparison models input the same bounding box sequence. On this basis, the inference weights of SPDR-Net are seamlessly integrated into the tracking logic. In addition, OSNet, which is widely adopted in the Multi-Object Tracking (MOT) field in the industry, is selected as the baseline ReID model for comparison. In the evaluation of continuous video streams, the core indicator is no longer the static mAP. Instead, it focuses on whether the identity trajectory of the same person breaks or jumps under cross-camera movement and complex interactions.

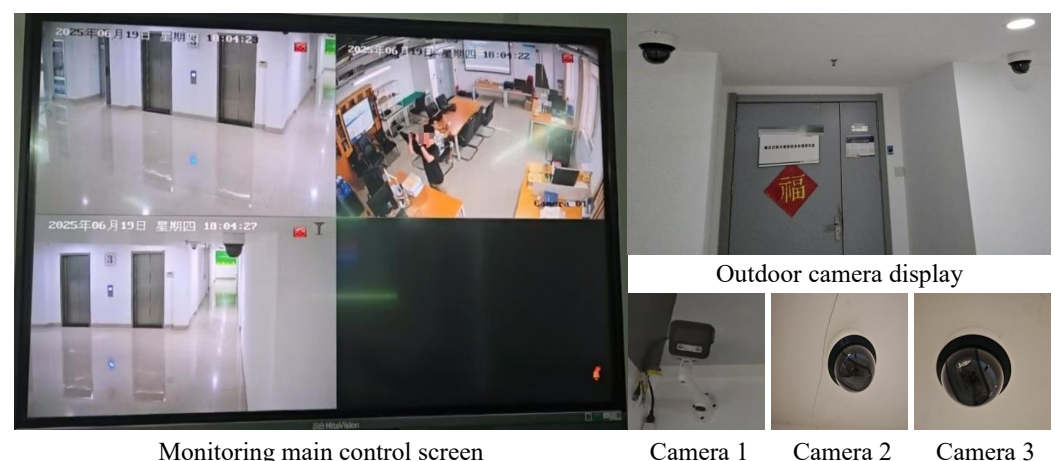


Figure 17. Multi-camera software and hardware testing system based on SPDR-Net. Any non-English text visible in the photographed environment is incidental background signage and is not used in the experiment.

In the continuous scene collection test, the tracking stability of the two models shows significant differences. The comparison of some keyframes from the test is shown in Figure 18 (OSNet baseline results) and Figure 19 (SPDR-Net test results).

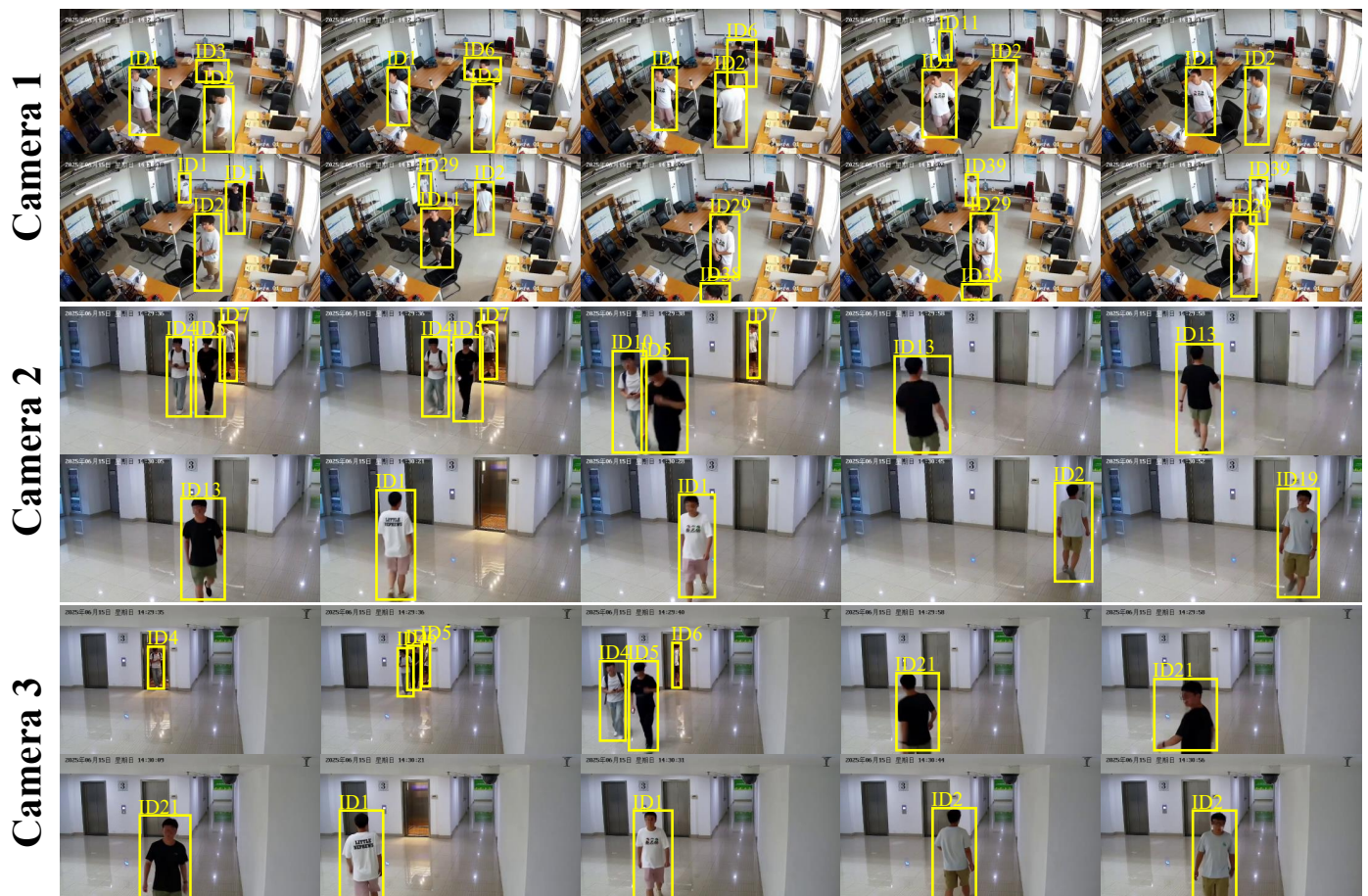


Figure 18. Person ReID and detection results of the baseline model (OSNet) under video streams (zoom-in is recommended). Any non-English text visible in the scene is incidental background signage and is unrelated to the analysis.

In the test of Figure 18, OSNet exhibits severe ID fragmentation. For example, when the target (initially ID3) moves indoors (Camera 1), its ID frequently breaks into ID6 and ID11 due to many indoor obstacles. When the target performs a “sitting down” action, the drastic deformation of the posture topology causes its ID to break again into ID38. In addition, when the target crosses indoor and outdoor cameras (entering Camera 2 and Camera 3), the global mutation of indoor/outdoor color temperature and brightness directly invalidates the original feature vector, triggering a systematic misidentification. This is because the traditional model heavily relies on static feature fusion at a global scale. Therefore, it cannot effectively deal with global domain shift situations in the open world.

In contrast, SPDR-Net in Figure 19 demonstrates excellent anti-interference ability and cross-camera identity consistency. Except for a brief ID confusion during extreme physical overlap (e.g., several frames where two persons completely occlude each other in Camera 1), SPDR-Net maintains a continuous lock on the target ID in the vast majority of complex scenes. This is because when the target is severely occluded or deformed in the lower body due to “sitting down”, the PGDR module quickly perceives the abnormality of the spatial topology. Using the sparse routing mechanism, it automatically down-weights the lower body features and relies predominantly on the head and torso experts to complete stable inter-frame matching. At the same time, when the target shuttles indoors and outdoors (Camera 1 to Camera 3), causing drastic changes in ambient light, the GL-Fusion module still firmly grasps the unchanged detailed features (such as clothing details and gait

contours) extracted by local experts, even when global color drifting occurs. This achieves a smooth transition of cross-domain features.

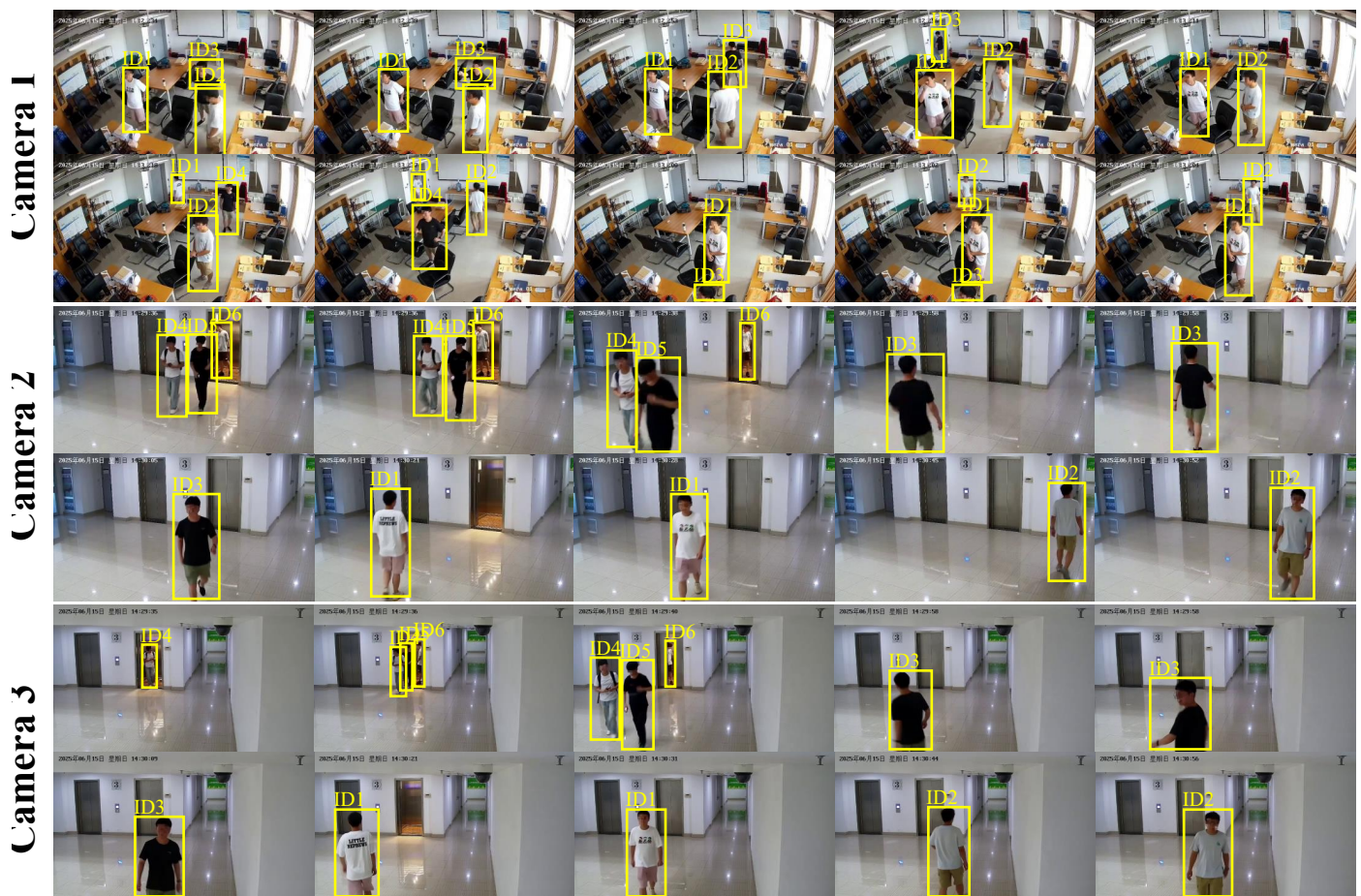


Figure 19. Person ReID and detection results of SPDR-Net under video streams (zoom-in is recommended). Any non-English text visible in the scene is incidental background signage and is unrelated to the analysis.

In summary, the closed-loop software and hardware test in this section not only confirms that ReID technology has high industrial deployment value. Moreover, it provides a robust dynamic routing architectural paradigm for ReID systems in the open world.

6. Discussion

SPDR-Net drives open-world ReID from static feature alignment to dynamic adaptive routing. Traditional models suffer from irreversible noise pollution under severe local occlusion. Based on the Mixture of Experts (MoE) model, SPDR-Net utilizes scene semantics to dynamically evaluate part confidence. Constrained by sparse entropy loss, the PGDR mechanism effectively isolates asymmetric interference and refines extracted features. Despite excellent performance, this spatial decoupling strategy faces limitations during full-element physical overlap (i.e., total occlusion by another person). In such extreme cases, single-frame part experts lose all visual cues, leading to brief identity confusion. In future work, we plan to extend SPDR-Net to a joint spatiotemporal dimension. By incorporating temporal contexts, the system will use unoccluded features from adjacent frames to compensate for severe current-frame occlusions. Additionally, model lightweighting techniques will be introduced to reduce inference latency on edge computing devices.

7. Conclusions

To address the problem of person ReID performance degradation caused by spatially asymmetric interference widely existing in open-world video surveillance systems, this paper proposes a novel Scene-Prompt-Driven Dynamic Routing Expert Network (SPDR-Net). This framework breaks through the limitations of traditional static feature fusion. It innovatively regards person part features as independent expert paths. Furthermore, it utilizes the Prompt-Guided Dynamic Routing (PGDR) network to execute dynamic weight routing and feature limitation driven by the scene context. To prevent the model from falling into trivial decisions, this paper introduces the sparse routing entropy loss to enhance the decisiveness of the network under extreme interference. Simultaneously, a global–local feature collaborative fusion module is designed to maintain the integrity of the macroscopic identity topology. To validate the effectiveness of the proposed framework, comprehensive experiments are conducted on seven mainstream public datasets and the newly constructed SD-ReID dataset (including Market1501, MSMT17, Occ-Duke, etc.). The results demonstrate that SPDR-Net exhibits outstanding scene adaptability under multi-scene domain shifts. Moreover, this paper constructs and open-sources a challenging real-world block dataset, SD-ReID. By building a multi-camera software and hardware closed-loop actual measurement system based on the YOLO detector, we verify the capability of the proposed architecture to maintain cross-camera identity consistency in real complex scenes. In summary, SPDR-Net not only provides a new dynamic routing paradigm for anti-interference feature representation at the theoretical level. More importantly, it provides a highly valuable experimental solution for highly robust security surveillance systems in engineering practice.

Author Contributions: Conceptualization, P.D. and H.L.; methodology, P.D. and H.L.; software, P.D. and B.W.; validation, P.D., X.G. and B.W.; formal analysis, H.L. and X.G.; investigation, P.D. and J.L.; resources, H.L. and J.L.; data curation, P.D. and Y.Z.; writing—original draft preparation, P.D.; writing—review and editing, H.L. and X.G.; visualization, P.D. and Y.Z.; supervision, H.L.; project administration, H.L.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shandong Province Natural Science Foundation under grant ZR2021QF077, the Key R&D Plan of Shandong Province (Innovation Ability Improvement Project of Small and Medium-sized Science and Technology Enterprises) under grant 2024TSGC0123.

Data Availability Statement: These data were derived from the following resources available in the public domain: Market1501: <https://www.kaggle.com/datasets/pengcw1/market-1501/data> (accessed on 19 July 2026). Msmt17: <https://aistudio.baidu.com/datasetdetail/166508> (accessed on 19 July 2026). Occ-Duke: <https://github.com/lightas/Occluded-DukeMTMC-Dataset> (accessed on 19 July 2026). SYSU-mm01: <https://www.isee-ai.cn/project/RGBIRReID.html> (accessed on 19 July 2026). Celeb-ReID: <https://github.com/Huang-3/Celeb-reID> (accessed on 19 July 2026). PRCC: <https://www.isee-ai.cn/~yangqize/clothing.html> (accessed on 19 July 2026). Mlr-CUHK03: https://pan.baidu.com/share/init?surl=hMQZq0LAPhII5RQ_EDDiFg&pwd=pami (accessed on 19 July 2026). SD-ReID: <https://github.com/dongpeng011/SD-ReID> (accessed on 19 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ding, J.; Linze, L.; Rongchang, L.; Cong, W.; Tianyang, X.; Wu, X.J. Robust person re-identification approach with deep learning and optimized feature extraction. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo Workshops, Niagara Falls, ON, Canada, 15–19 July 2024; pp. 1–6. [CrossRef]
2. Cui, Z.; Zhou, J.; Peng, Y. DMA: Dual modality-aware alignment for visible-infrared person re-identification. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 2696–2708. [CrossRef]

3. Yin, Q.; Ding, G.; Zhang, T.; Gong, Y. Robust fine-grained learning for cloth-changing person re-identification. *Mathematics* **2025**, *13*, 429. [\[CrossRef\]](#)
4. Wang, Q.; Qian, X.; Li, B.; Xue, X.; Fu, Y. Exploring fine-grained representation and recombination for cloth-changing person re-identification. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 6280–6292. [\[CrossRef\]](#)
5. Zhang, F.; Ma, H.; Zhu, J.; Hamdulla, A.; Zhu, B. Frce: Transformer-based feature reconstruction and cross-enhancement for occluded person re-identification. *Expert Syst. Appl.* **2024**, *258*, 125110. [\[CrossRef\]](#)
6. Li, X.; Chen, Y.; Wu, Y.; Liang, Y.; Cao, Y.; Liu, Q.; Yuan, C. Occluded person re-identification method based on pedestrian background decoupling transformer. *Mathematics* **2026**, *14*, 1725. [\[CrossRef\]](#)
7. Ye, M.; Chen, S.; Li, C.; Zheng, W.S.; Crandall, D.; Du, B. Transformer for object re-identification: A survey. *Int. J. Comput. Vis.* **2025**, *133*, 2410–2440. [\[CrossRef\]](#)
8. Zheng, W.S.; Yan, J.; Peng, Y.X. A versatile framework for multi-scene person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 1362–1380. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Li, S.; Sun, L.; Li, Q. CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 1405–1413. [\[CrossRef\]](#)
10. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [\[CrossRef\]](#)
11. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496. [\[CrossRef\]](#)
12. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM International Conference on Multimedia (MM '18), Seoul, Republic of Korea, 22–26 October 2018; pp. 274–282. [\[CrossRef\]](#)
13. Zhou, L.; Li, S.; Dong, N.; Tai, Y.; Zhang, Y.; Li, H. Hierarchical prompt learning for image- and text-based person re-identification. *arXiv* **2025**, arXiv:2511.13575. [\[CrossRef\]](#)
14. Zheng, W.; Gong, S.; Xiang, T. Towards open-world person re-identification by one-shot group-based verification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 591–606. [\[CrossRef\]](#) [\[PubMed\]](#)
15. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. TransReID: Transformer-based object re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 15013–15022. [\[CrossRef\]](#)
16. Chen, Z.; Dong, Q.; Gong, S.; Zhu, X. Inter-task association critic for cross-resolution person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 2605–2615. [\[CrossRef\]](#)
17. Gao, Z.; Wei, S.; Guan, W.; Zhu, L.; Wang, M.; Chen, S. Identity-guided collaborative learning for cloth-changing person reidentification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 2819–2837. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Wen, W.; Chen, H.; Zhang, S. Prompt-Anchored Vision-Text Distillation for Lifelong Person Re-Identification. *arXiv* **2026**, arXiv:2605.05027. [\[CrossRef\]](#)
19. Wu, R.; Jiao, B.; Wang, W.; Liu, M.; Wang, P. Enhancing visible-infrared person re-identification with modality- and instance-aware visual prompt learning. In Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), Phuket, Thailand, 10–14 June 2024; pp. 579–588. [\[CrossRef\]](#)
20. Jiang, X.; Yu, L.; Lin, G.; Zhang, Y. PMD-Transformer: A domain generalization approach for person re-identification. *IEEE Access* **2025**, *13*, 78921–78931. [\[CrossRef\]](#)
21. Lyu, H.; Yao, H.; Xu, C. Multiple local prompts distillation for domain generalization. *IEEE Trans. Multimed.* **2025**, *27*, 8555–8564. [\[CrossRef\]](#)
22. Wangning; Xie, L.; Lu, S.; Gan, S. PVTNL: Prompting vision transformers with natural language for generalizable person re-identification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, 4–9 November 2025 ; pp. 21663–21674. [\[CrossRef\]](#)
23. Xu, X.; Li, W.; Liu, W.; Huang, W.; Yu, Z.; Yang, B.; Liao, X.; Jiang, K. FedBPrompt: Federated domain generalization person re-identification via body distribution aware visual prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 3–7 June 2026; pp. 40448–40457. Available online: https://openaccess.thecvf.com/content/CVPR2026/html/Xu_FedBPrompt_Federated_Domain_Generalization_Person_Re-Identification_via_Body_Distribution_Aware_CVPR_2026_paper.html (accessed on 19 July 2026).
24. Zeng, Z.; Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.Y.; Satoh, S. Illumination-adaptive person re-identification. *IEEE Trans. Multimed.* **2020**, *22*, 3064–3074. [\[CrossRef\]](#)

25. Wang, Z.; Zhu, F.; Tang, S.; Zhao, R.; He, L.; Song, J. Feature erasing and diffusion network for occluded person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 4754–4763. [\[CrossRef\]](#)
26. Mishra, R.K.; Mondal, A.; Mathew, J. Swin transformer with part-level tokenization for occluded person re-identification. *Mach. Vis. Appl.* **2025**, *36*, 1–17. [\[CrossRef\]](#)
27. Zheng, W.; Li, X.; Xiang, T.; Liao, S.; Lai, J.; Gong, S. Partial person re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4678–4686. [\[CrossRef\]](#)
28. Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Pinto, A.S.; Keysers, D.; Houlsby, N. Scaling vision with sparse mixture of experts. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2021**, *34*, 8583–8595. Available online: <https://openreview.net/forum?id=FrIDgjDOH1u> (accessed on 19 July 2026).
29. Puigcerver, J.; Riquelme, C.; Mustafa, B.; Houlsby, N. From sparse to soft mixtures of experts. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024. [\[CrossRef\]](#)
30. Zhu, L.; Chen, T.; Ji, D.; Ye, J.; Liu, J. Not every patch is needed: Toward a more efficient and effective backbone for video-based person re-identification. *IEEE Trans. Image Process.* **2025**, *34*, 785–800. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Li, Y.; Song, L.; Chen, Y.; Li, Z.; Zhang, X.; Wang, X.; Sun, J. Learning dynamic routing for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8553–8562. [\[CrossRef\]](#)
32. Somers, V.; Alahi, A.; De Vleeschouwer, C. Keypoint promptable re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 1–18. [\[CrossRef\]](#)
33. Wang, Z.; Li, C.; Li, P.; Zheng, A.; Tang, J.; Luo, B. Prototype-based diversity and integrity learning for all-day multi-modal person re-identification. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 11385–11400. [\[CrossRef\]](#)
34. Sun, Z.; Tan, L.; Shen, Y.; Cai, C.; Sun, X.; Dai, P.; Cao, L.; Ji, R. FlexiReID: Adaptive mixture of expert for multi-modal person re-identification. In Proceedings of the International Conference on Machine Learning (ICML), Vancouver, BC, Canada, 13–19 July 2025. [\[CrossRef\]](#)
35. Wang, Y.; Liu, Y.; Zheng, A.; Zhang, P. DeMo: Decoupled feature-based mixture of experts for multi-modal object re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 8141–8149. [\[CrossRef\]](#)
36. Dong, W.; Yang, X.; Cheng, D.; Wang, N.; Gao, X. Escaping modal interactions: An efficient DESANet for multi-modal object re-identification. *IEEE Trans. Image Process.* **2025**, *34*, 5068–5083. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1116–1124. [\[CrossRef\]](#)
38. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person re-identification in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8476–8484. [\[CrossRef\]](#)
39. Miao, J.; Wu, Y.; Dai, P.; Li, Y.; Ji, R.; Yang, Y. Pose-guided feature alignment for occluded person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, 27 October–2 November 2019; pp. 542–551. [\[CrossRef\]](#)
40. Jiao, J.; Zheng, W.S.; Wu, A.; Zhu, X.; Gong, S. Deep low-resolution person re-identification. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, USA, 2–7 February 2018; pp. 6967–6974. [\[CrossRef\]](#)
41. Wu, A.; Zheng, W.S.; Yu, H.H.; Gong, S.; Lai, J. RGB-infrared cross-modality person re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5380–5389. [\[CrossRef\]](#)
42. Huang, Y.; Xu, J.; Wu, Q.; Zhong, Y.; Zhang, P.; Zhang, Z. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 3459–3471. [\[CrossRef\]](#)
43. Yang, Q.; Wu, A.; Zheng, W.S. Person re-identification by contour sketch under moderate clothing change. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2029–2046. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Dong, P.; Liu, H.; Guo, X.; Li, J.; Zhou, Y.; Wang, B. SD-ReID Dataset and SPDR-Net Source Code, Version 1.0. GitHub Repository. 2026. Available online: <https://github.com/dongpeng011/SD-ReID> (accessed on 19 July 2026).
45. Dong, N.; Zhang, L.; Yan, S.; Tang, H.; Tang, J. Erasing, transforming, and noising defense network for occluded person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 4458–4470. [\[CrossRef\]](#)
46. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021. [\[CrossRef\]](#)
47. Zhu, K.; Guo, H.; Yan, T.; Zhu, Y.; Wang, J.; Tang, M. PASS: Part-aware self-supervised pre-training for person re-identification. In *Computer Vision—ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; pp. 198–214. [\[CrossRef\]](#)

48. Chen, W.; Xu, X.; Jia, J.; Luo, H.; Wang, Y.; Wang, F.; Jin, R.; Sun, X. Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 18–22 June 2023; IEEE Computer Society: Los Alamitos, CA, USA, 2023; pp. 15050–15061. [[CrossRef](#)]
49. Hu, B.; Wang, X.; Liu, W. Personvit: Large-scale self-supervised vision transformer for person re-identification. *Mach. Vis. Appl.* **2025**, *36*, 32. [[CrossRef](#)]
50. Yuan, C.; Zhang, G.; Ma, C.; Zhang, T.; Niu, G. From poses to identity: Training-free person re-identification via feature centralization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 11–15 June 2025; pp. 24409–24418. Available online: https://openaccess.thecvf.com/content/CVPR2025/html/Yuan_From_Poses_to_Identity_Training-Free_Person_Re-Identification_via_Feature_Centralization_CVPR_2025_paper.html (accessed on 19 July 2026).
51. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3701–3711. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.