

Semantic-Conditioned Variational Compensation and Uncertainty-Guided Retrieval for Occluded Person Re-Identification^{*,**}

Xiaohao Xie^{a,1}, Wenhua Jiao^{a,*}

ARTICLE INFO

Keywords:

Occluded person re-identification
Vision-language models
Masked feature compensation
Variational inference
Uncertainty-guided retrieval

ABSTRACT

Person re-identification (ReID) under severe occlusion remains a formidable challenge, as incomplete observations inevitably lead to fragmented features and corrupted distance metrics. While recent vision-language methods leverage text prompts to passively refine observed representations, they are prone to spatial drift under heavy occlusion and struggle to explicitly recover missing topological structures. Furthermore, existing feature recovery pipelines often rely on deterministic imputation, blindly fusing hallucinated features into the metric space without considering the inherent uncertainty of guessing. To break this bottleneck, we propose a unified framework that seamlessly integrates semantic probing with uncertainty-guided feature compensation. Specifically, we first introduce Text-Driven Proxy Pooling (TDPP) to actively retrieve reliable body-part evidence from corrupted observations using text semantics as spatial anchors. Subsequently, we propose Prompt-Guided Masked Feature Compensation (PMFC), which formulates occlusion recovery as a conditional variational masked modeling problem. By jointly optimizing reconstruction and KL divergence, PMFC explicitly encodes the recovery ambiguity into a posterior variance. During inference, this generative variance acts as a natural safety shield through our Uncertainty-guided Dynamic Retrieval (UDR) mechanism, strictly penalizing unreliable occluded features and preventing hallucinated noise from contaminating the metric space. Extensive experiments demonstrate that our framework achieves state-of-the-art (SOTA) or comparable SOTA performance across both occluded and holistic benchmarks (e.g., yielding 93.4% Rank-1 and 91.9% mAP on Occluded-ReID).

1. Introduction

Person re-identification (ReID) aims to retrieve images of the same individual across multiple non-overlapping camera views and has become a fundamental technique in intelligent surveillance, pedestrian retrieval, and public security applications. Although remarkable progress has been achieved under holistic observation settings, real-world deployment remains challenging because pedestrians are frequently occluded by surrounding people or objects. Occlusion simultaneously removes identity-discriminative information, introduces distracting visual patterns, and disrupts the spatial continuity of body parts, making reliable representation learning significantly more difficult.

To address these challenges, recent studies have progressively improved occluded person ReID through structural modeling [11, 7, 17], feature recovery [21, 23], relational reasoning [8, 19], and, more recently, vision-language semantic guidance [4, 1, 18, 15]. Inspired by foundational vision-language pre-training [13, 6], introducing linguistic knowledge into person ReID has demonstrated stronger

semantic alignment and richer prior knowledge than conventional appearance-based learning [10]. Meanwhile, recent surveys on occluded person ReID indicate that future progress increasingly relies on combining semantic understanding with robust reasoning under severe information loss rather than solely improving feature extraction pipelines [12]. Despite these advances, accurately identifying reliable body-part evidence and effectively compensating for missing identity information under heavy occlusion remain open challenges.

Existing occluded ReID methods generally follow three representative paradigms. The first paradigm improves feature extraction by introducing pose priors, part discovery, or relational modeling to obtain more robust local representations [11, 7, 2, 8, 19]. The second paradigm exploits vision-language semantics through prompt learning or semantic alignment to enhance representation quality and body-part consistency [4, 1, 18, 15]. The third paradigm treats occlusion as a feature recovery problem and reconstructs incomplete representations using contextual information or cross-image correspondences [21, 23, 8]. These approaches have substantially improved robustness against partial occlusion and together constitute the current mainstream solutions for occluded person ReID.

Nevertheless, directly deploying these state-of-the-art paradigms under severe occlusion exposes a fundamental dilemma between **cross-modal localization** and **reliable feature compensation**. First, cutting-edge vision-language models (e.g., ProFD [1], COPE [15]) predominantly utilize

^{*}This document is the results of the research project funded by the National Science Foundation.

^{**}The second title footnote which is a longer text matter to fill through the whole text width and overflow into another line in the footnotes area of the first page.

 xh_xie@cumt.edu.cn (X. Xie); wjiao@cumt.edu.cn (W. Jiao)
 https://faculty.cumt.edu.cn/JWH/zh_CN/index.htm (W. Jiao)
ORCID(s):

text prompts passively as post-hoc alignment signals or representation refiners for observed regions. Without explicit spatial-geometric regularization, their cross-attention mechanisms easily suffer from semantic drift toward background clutter when target body parts are heavily occluded, and they lack a generative mechanism to explicitly reconstruct missing topological structures. Second, while advanced feature recovery networks (e.g., FED [21], FRT [23]) attempt to impute missing representations, they largely rely on **deterministic imputation** conditioned on visual context or cross-image correspondences. Because recovering completely unseen regions is an inherently ill-posed problem, deterministic guessing inevitably produces feature hallucinations. Blindly fusing these hallucinated representations into the metric space with equal confidence severely contaminates the identity distance calculation during retrieval.

To resolve this fundamental dilemma, we argue that an ideal occluded ReID framework must transition from *passive alignment and deterministic guessing* to *active semantic probing and uncertainty-aware variational compensation*. Built upon the strong ProFD baseline [1], we propose a unified pipeline that tightly couples vision-language priors with probabilistic generative modeling. Specifically, to prevent semantic drift, we first design **Text-Driven Proxy Pooling (TDPP)**, which transforms body-part text embeddings into active spatial queries to interrogate visual patches under a decoupled visibility-aware spatial distillation constraint. Subsequently, to rescue missing topologies without deterministic bias, we propose **Prompt-Guided Masked Feature Compensation (PMFC)**. By physically filling occluded visual slots with explicit text prompts, PMFC formulates feature recovery as a conditional variational masked modeling problem. Through the joint optimization of visible-part reconstruction and semantic-prior Kullback-Leibler (KL) divergence, the network explicitly encodes recovery ambiguity into its latent posterior variance. Crucially, during inference, this generative variance unlocks our **Uncertainty-guided Dynamic Retrieval (UDR)** mechanism, acting as a natural confidence shield that dynamically down-weights ambiguous features and safely isolates hallucinated noise from the retrieval metric space.

The main contributions of this work are summarized as follows:

- We propose Text-Driven Proxy Pooling (TDPP), which uses body-part semantic prompts as query-like proxies to dynamically retrieve reliable local evidence from corrupted visual representations. A visibility-aware spatial distillation objective further regularizes semantic localization under occlusion.
- We propose Prompt-Guided Masked Feature Compensation (PMFC), which formulates occluded-part recovery as a semantically conditioned variational masked modeling problem.
- By jointly optimizing the reconstruction and KL divergence losses, PMFC naturally leverages its posterior variance as a spatial confidence prior during

inference, enabling uncertainty-guided dynamic retrieval that safely isolates hallucinated noise from the distance metric.

2. Related Work

We review the literature most relevant to our work from three perspectives: occluded person re-identification, vision-language modeling for ReID, and masked feature modeling and compensation. Rather than exhaustively surveying person ReID, we focus on approaches related to the challenges addressed in this work: reliable part localization under corrupted observations, inference of missing identity cues, and uncertainty-aware retrieval.

2.1. Occluded Person Re-Identification

Occluded person ReID is challenging because occlusion simultaneously removes identity-discriminative cues and introduces interference from obstacles or surrounding pedestrians. Earlier approaches alleviate these problems by introducing explicit structural cues. PGFA employs pose landmarks to identify visible body regions and suppress occlusion noise [11], while PAT uses a Transformer encoder-decoder architecture to discover discriminative body parts without relying on rigid spatial partitions [7]. Subsequent methods increasingly focus on adaptive suppression and recovery of corrupted representations. FED handles both non-pedestrian occlusions and interfering pedestrians through feature erasing and diffusion [21], whereas FRT combines visibility-aware matching with Transformer-based feature recovery to complement incomplete query representations [23]. Recent methods further explore semantic region generation [4], realistic occlusion simulation [17], and multi-level relational modeling [8, 19].

Alongside these appearance-driven approaches, semantic guidance has become increasingly important. RGA Net leverages CLIP-derived semantic prototypes and learnable prompts to generate and assess human body regions [4], while ProFD employs part-specific prompts to facilitate spatially and semantically consistent part-feature disentanglement [1]. These studies substantially improve robustness under occlusion, but semantic priors are mainly used for region localization, feature alignment, or representation refinement. How to use explicit part semantics as query-like probes for searching reliable evidence under severely corrupted observations remains less explored, which motivates the semantic probing perspective adopted in this work.

2.2. Vision-Language Models in Re-ID

Large-scale vision-language pre-training provides a promising means of introducing semantic knowledge beyond conventional appearance supervision. CLIP learns transferable image-text representations through large-scale contrastive pre-training [13]. Building upon this paradigm, CLIP-ReID adapts CLIP to image ReID without requiring manually annotated text descriptions by learning identity-specific textual tokens and using the resulting textual representations to guide visual representation learning [6].

More recent studies move from global vision-language transfer toward finer-grained semantic modeling. ProFD introduces part-specific prompts for occluded ReID and aligns textual semantics with spatial body information to disentangle local representations [1]. THCB-Net further incorporates textual information into hierarchical visual context modeling for occluded person ReID [18]. Most recently, COPE combines occlusion-consistency learning with prompt enhancement, further demonstrating the potential of prompt-based semantic modeling for handling occluded observations [15].

Despite this progress, language is predominantly exploited as semantic supervision, alignment guidance, or prompt-based representation refinement [1, 6, 18, 15]. The complementary role of language as an explicit probing signal—i.e., using body-part semantics to interrogate spatial representations for reliable identity evidence—has received considerably less attention. This distinction is central to our TDPP design: the semantic prior is not merely aligned with an already extracted representation, but is used to guide part-specific evidence discovery under visual corruption.

2.3. Masked Feature Modeling and Compensation

Occlusion can also be viewed as a structured missing-information problem. Masked image modeling demonstrates that missing visual content can be inferred from contextual observations. Representative approaches such as MAE [3] and SimMIM [22] learn visual representations by reconstructing masked image content, establishing contextual prediction as an effective representation-learning paradigm.

Related ideas have emerged in occluded ReID at the feature level. FED redistributes useful pedestrian information after suppressing contaminated features [21], while FRT recovers incomplete query representations by exploiting feature sets from reliable nearest-neighbor gallery samples selected through visibility-aware matching [23]. Other methods exploit semantic region estimation or multi-level feature relations to preserve discriminative information when local observations become unreliable [4, 8, 19]. These approaches demonstrate that simply removing corrupted regions is often insufficient; missing identity cues must also be compensated for.

However, existing recovery mechanisms are mainly conditioned on visual context, cross-image correspondences, learned prototypes, or feature relations [21, 23, 8, 19]. Even when semantic prompts are introduced for localization or feature disentanglement [4, 1], they are not explicitly coupled with a masked compensation mechanism that specifies which anatomical semantics should be recovered. This motivates a semantically conditioned formulation in which part-level prompts specify the expected semantic content while reliable visual context provides identity-specific evidence for feature compensation.

Overall, existing studies have made substantial progress in occlusion-aware representation learning, semantic prompting, and feature recovery. However, three coupled issues remain insufficiently addressed: semantic priors are mainly used to refine observed features rather than actively probe

reliable part evidence; missing-feature recovery is rarely conditioned by explicit part semantics; and the structural uncertainty of compensated features is rarely utilized to penalize unreliable distance calculations during retrieval. These limitations motivate our TDPP, PMFC, and UDR mechanisms, respectively.

3. Proposed Method

Overview. The proposed framework builds upon the strong ProFD baseline [1] to tackle severe feature misalignment and topological information loss caused by occlusion in person re-identification. As illustrated in Figure 1, our pipeline consists of two complementary representation modules coupled with an uncertainty-aware inference mechanism. First, given the visual patch sequence V and body-part text embeddings T , **Text-Driven Proxy Pooling (TDPP)** (Sec. 3.1) actively probes and aggregates reliable local visual evidence F_{part} by utilizing semantic text priors as strict spatial anchors under decoupled spatial supervision. Second, to recover missing semantics in occluded regions, we introduce **Prompt-Guided Masked Feature Compensation (PMFC)** (Sec. 3.2), which formulates occlusion recovery as a conditional variational masked modeling problem. By blending corrupted visual features with explicit text prompts and regularizing the latent space via joint self-reconstruction ($\mathcal{L}_{\text{recon}}$) and semantic-prior KL divergence ($\mathcal{L}_{\text{KL}}$), PMFC explicitly encodes recovery ambiguity into its posterior variance. Finally, during the testing phase, this generative variance naturally unlocks **Uncertainty-guided Dynamic Retrieval (UDR)**, which derives a structural confidence score C_{vae} to re-weight part visibility S_{DAI} and securely replace severely corrupted features with compensated representations F_{comp} .

3.1. Text-Driven Proxy Pooling (TDPP)

Existing part-based methods typically rely on 2D heuristic spatial pooling, which is easily corrupted when occlusions disrupt visual continuity. To address this, we introduce TDPP, which leverages explicit text prompts to dynamically address and group visual patches.

Given a flattened visual patch sequence $V \in \mathbb{R}^{L \times d}$ and the text embeddings of K body parts $T \in \mathbb{R}^{K \times d}$, we employ a Transformer Decoder for cross-attention. A crucial design is the residual prompt injection mechanism, where we directly feed the text embeddings T as the initial target sequence:

$$F_{\text{part}}, A_{\text{spatial}} = \text{Decoder}(\text{tgt} = T, \text{memory} = V, \text{query_pos} = T). \quad (1)$$

The output local features can be formulated as $F_{\text{part}} \approx T + \text{CrossAttn}(T, V)$. This formulation ensures that text semantics serve as a robust anchor under severe occlusion, while seamlessly preserving fine-grained visual details in unoccluded scenarios.

Decoupled Optimization with Spatial Distillation. To optimize the proxy module efficiently without introducing categorical ambiguity, we decouple the raw attention map

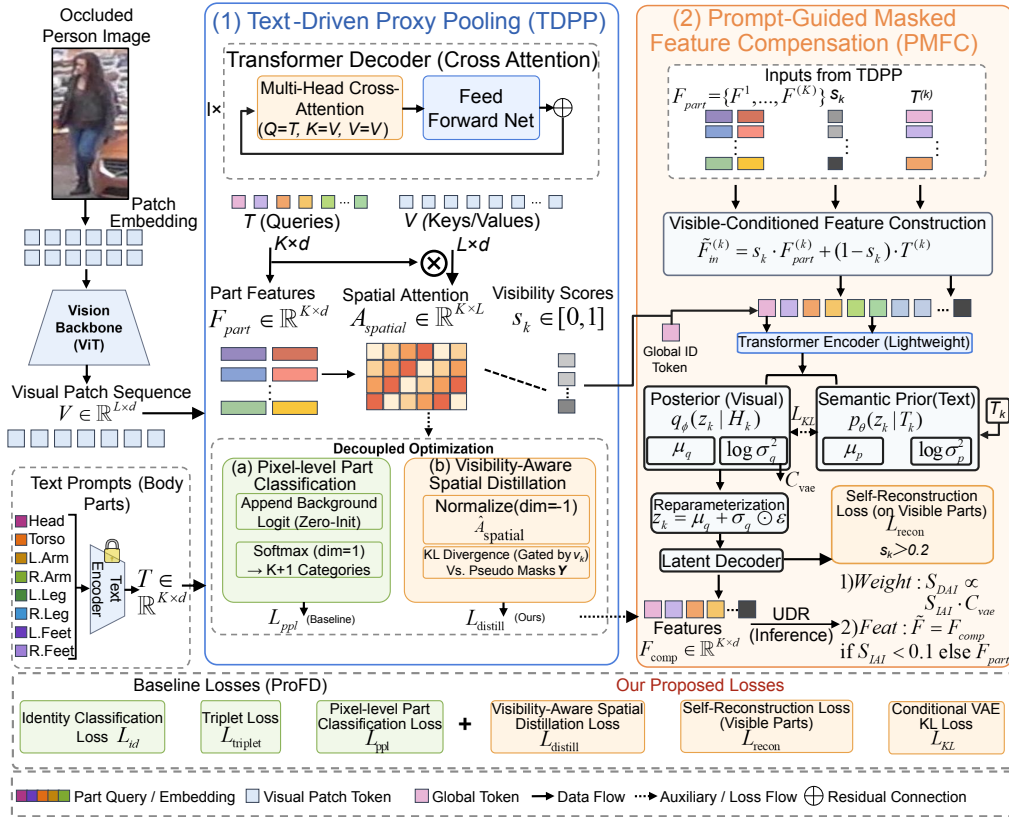


Figure 1: Overview of the proposed framework for occluded person re-identification. Given an occluded person image, the vision backbone extracts the visual patch sequence $V \in \mathbb{R}^{L \times d}$, while the frozen text encoder encodes $K = 8$ anatomical body-part prompts into semantic embeddings $T \in \mathbb{R}^{K \times d}$. **(1) Text-Driven Proxy Pooling (TDPP)** employs T as explicit spatial queries in a Transformer Decoder to aggregate part features F_{part} and spatial attention maps $A_{spatial}$, regulated by a decoupled optimization strategy combining baseline pixel-level part classification (L_{ppl}) and visibility-aware spatial distillation ($L_{distill}$). **(2) Prompt-Guided Masked Feature Compensation (PMFC)** constructs visibility-conditioned tokens $\tilde{F}_{in}^{(k)}$ via cross-modal filling and infers the latent visual posterior $q_\phi(z_k | H_k)$. Guided by a semantic conditional prior $p_\theta(z_k | T_k)$ via L_{KL} and a visible-part self-reconstruction objective L_{recon} , the latent decoder generates compensated representations F_{comp} . During inference, **Uncertainty-guided Dynamic Retrieval (UDR)** extracts the posterior variance $\log \sigma_q^2$ as a structural confidence prior C_{vae} to dynamically modulate retrieval weights S_{DAI} and safely substitute severely occluded features ($S_{DAI} < 0.1$).

$A_{spatial} \in \mathbb{R}^{K \times L}$ into two distinct manifolds. For pixel-level part classification, we must maintain a strict mutually exclusive competition among different body parts and the background. Following the baseline paradigm, we append a zero-initialized background logit to $A_{spatial}$ and perform channel-wise Softmax normalization (dim=1). This yields a smooth probability map across $K + 1$ categories, which is directly supervised by the baseline’s pixel-level part classification loss (L_{ppl}). This explicit categorical competition ensures clean semantic boundaries.

For spatial localization, we introduce a Visibility-Aware Spatial Distillation loss. During training, we utilize an off-the-shelf human parsing model (e.g., Mask R-CNN) to generate pseudo-parsed target masks $Y \in \mathbb{R}^{K \times L}$, where $Y^{(k)}$ acts as the spatial ground-truth for the k -th body part. We normalize $A_{spatial}$ along the spatial dimension (dim=-1) to obtain $\hat{A}_{spatial}$, aligning it with Y . To prevent noisy gradients from occluded regions, the loss is gated by a binary visibility indicator $v_k \in \{0, 1\}$, which is set to 1 if the k -th part is

present in the pseudo-mask ($\sum Y^{(k)} > 0$):

$$\mathcal{L}_{distill} = \frac{1}{\sum_{k=1}^K v_k} \sum_{k=1}^K v_k \cdot D_{KL}(Y^{(k)} \parallel \hat{A}_{spatial}^{(k)}), \quad (2)$$

By strictly aligning the probability distributions within the spatial dimension, $\mathcal{L}_{distill}$ explicitly forces the proxy queries to attend to geometrically correct regions.

3.2. Prompt-Guided Masked Feature Compensation (PMFC)

Occlusion inevitably leads to missing local semantics. Since recovering unseen regions is highly uncertain, we formulate occlusion recovery as a feature-level conditional variational masked modeling problem.

Conditional Variational Recovery. Let $F_{part} \in \mathbb{R}^{K \times d}$ denote the visual part features extracted by TDPP. Unlike conventional masked autoencoders that employ generic learnable mask tokens to impute missing regions, we propose a cross-modal filling strategy. For each part k , we utilize

the continuous visibility score $s_k \in [0, 1]$ (derived from the spatial maximum of the baseline's attention maps) to softly blend the visual feature and its corresponding text semantic prompt $T^{(k)}$. The text prompt serves as an explicit semantic prior to physically fill the occluded visual space:

$$\tilde{F}_{in}^{(k)} = s_k \cdot F_{part}^{(k)} + (1 - s_k) \cdot T^{(k)} \quad (3)$$

To provide a holistic identity context, we prepend a global identity feature (extracted via global average pooling from the visual backbone) to the sequence as a prompt token. The augmented sequence is then processed by a lightweight Transformer Encoder to obtain contextualized part representations H_k , from which we infer the latent posterior distribution $q_\phi(z_k|H_k)$, parameterized by a mean μ_q and a log-variance $\log \sigma_q^2$. After reparameterization ($z_k = \mu_q + \sigma_q \odot \epsilon$ during training, while directly utilizing the posterior mean μ_q during inference for deterministic evaluation), the latent vectors for the K body parts are decoded via a residual Multi-Layer Perceptron (MLP) to yield the compensated features $F_{comp} \in \mathbb{R}^{K \times d}$.

To ensure the compensator learns meaningful structural representations without introducing arbitrary feature distortions, we apply a visibility-aware self-reconstruction loss strictly on the valid visible parts. Instead of applying a rigid hard threshold, we filter out extremely unreliable regions with a low confidence threshold ($\tau_v = 0.2$) and utilize the continuous visibility scores s_k as adaptive reconstruction weights. This encourages F_{comp} to robustly reconstruct the reliable F_{part} representations with a focus on highly visible body parts:

$$\mathcal{L}_{recon} = \frac{1}{\sum_{k=1}^K \mathbb{I}(s_k > \tau_v) s_k} \sum_{k=1}^K \mathbb{I}(s_k > \tau_v) s_k \cdot \left[\mathcal{L}_{cos}(F_{comp}^{(k)}, F_{part}^{(k)}) + \alpha \mathcal{L}_{smooth}(F_{comp}^{(k)}, F_{part}^{(k)}) \right], \quad (4)$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathcal{L}_{cos}(\cdot, \cdot)$ denotes the cosine distance, and $\mathcal{L}_{smooth}(\cdot, \cdot)$ is the Smooth-L1 loss with a scaling factor $\alpha = 0.1$ to prevent feature scale drift. Meanwhile, to prevent identity collapse under severe occlusion, the latent posterior is regularized by a conditional Kullback-Leibler (KL) divergence loss $\mathcal{L}_{KL}$. Instead of using a standard isotropic Gaussian, we define a semantic-conditioned prior $p_\theta(z_k|T^{(k)}) = \mathcal{N}(\mu_p, \sigma_p^2)$ parameterized by a prior encoder taking the text embedding $T^{(k)}$ as input. The loss explicitly aligns the visual posterior $q_\phi(z_k|H_k)$ with this semantic prior:

$$\mathcal{L}_{KL} = \frac{1}{K} \sum_{k=1}^K D_{KL}\left(q_\phi(z_k|H_k) \parallel p_\theta(z_k|T^{(k)})\right). \quad (5)$$

Uncertainty-guided Dynamic Retrieval (UDR). A critical flaw in standard deterministic feature recovery pipelines is that generated representations are directly fused into the distance metric during inference without evaluating their

reliability. Formulating PMFC as a variational autoencoder naturally resolves this issue. The predicted log-variance natively reflects the network's internal generative uncertainty. During the inference stage, we extract this uncertainty $U_k = \frac{1}{d} \sum_{i=1}^d \exp(\log \sigma_{k,i}^2)$ for each part k . To maximize the contrast between reliable and corrupted regions, we apply a min-max normalization across the K parts within the same instance and map it to a structural confidence score:

$$C_{vae}^{(k)} = 1 - \frac{U_k - \min(U)}{\max(U) - \min(U) + \epsilon} \quad (6)$$

where ϵ is a small constant for numerical stability, and U denotes the uncertainty set of all K parts for the given instance. Simultaneously, the baseline extracts an Implicit Attention Inference (IAI) score, $S_{IAI}^{(k)} = \sigma(\text{Linear}(F_{part}^{(k)}))$, representing the inherent discriminative quality of the raw visual features. We perform a multiplicative fusion of these two orthogonal metrics:

$$S_{fused}^{(k)} = S_{IAI}^{(k)} \cdot C_{vae}^{(k)} \quad (7)$$

The fused score is then sharpened via a temperature-scaled Softmax function ($\tau_s = 0.1$) across the K body parts to produce the final Dynamic Attention Inference (DAI) visibility weights:

$$S_{DAI}^{(k)} = \frac{\exp(S_{fused}^{(k)}/\tau_s)}{\sum_{j=1}^K \exp(S_{fused}^{(j)}/\tau_s)} \quad (8)$$

This formulation guarantees a competitive weighting mechanism within each instance, where only features possessing both high discriminative quality and high generative certainty are assigned substantial weights during the distance calculation. Furthermore, if the raw visual discriminative score is critically low ($S_{IAI}^{(k)} < 0.1$), indicating severe occlusion, we securely replace the raw feature $F_{part}^{(k)}$ with the semantically generated feature $F_{comp}^{(k)}$ to salvage lost topologies. Finally, following the baseline retrieval protocol [1], the overall matching distance between a query and a gallery instance is computed by combining the global feature cosine distance with the local part-to-part distances weighted by the mutual $S_{DAI}^{(k)}$ scores. It is worth noting that UDR is intrinsically conditioned on the joint optimization of both $\mathcal{L}_{recon}$ and $\mathcal{L}_{KL}$. In our ablation studies, if either loss is disabled, the generative variance loses its probabilistic significance, and the framework safely degenerates to the baseline inference logic to prevent noise contamination.

Overall Objective. The entire framework is trained end-to-end. To ensure a fair and robust foundation, we inherit the base optimization objectives from the ProFD baseline, including the identity classification loss $\mathcal{L}_{id}$, the triplet loss $\mathcal{L}_{triplet}$, and the pixel-level part classification loss $\mathcal{L}_{ppl}$. The total loss combines these baseline losses with our newly proposed objectives:

$$\mathcal{L}_{total} = (\mathcal{L}_{id} + \mathcal{L}_{triplet} + \mathcal{L}_{ppl}) + \lambda_{distill} \mathcal{L}_{distill} + \lambda_{recon} \mathcal{L}_{recon} + \lambda_{KL} \mathcal{L}_{KL}, \quad (9)$$

where λ_{distill} , λ_{recon} , λ_{KL} are constant scaling hyperparameters.

4. Experiments

4.1. Datasets and Evaluation Protocols

We comprehensively evaluate our proposed framework on four widely adopted person ReID benchmarks, including two occlusion-specific datasets and two holistic datasets:

- **Occluded-Duke [11]:** The most challenging and extensively benchmarked occluded dataset, containing 15,618 training images, 2,210 occluded query images, and 17,661 gallery images.
- **Occluded-ReID [26]:** A dataset captured by mobile cameras specifically for occluded scenarios, comprising 2,000 images of 200 occluded identities.
- **Market-1501 [25]:** A classic holistic ReID dataset containing 32,668 images of 1,501 identities captured across six distinct cameras.
- **DukeMTMC-ReID [14]:** A large-scale holistic dataset comprising 36,411 images of 1,812 identities captured by eight high-resolution cameras.

Evaluation Protocols. Following standard practices, we adopt the Cumulative Matching Characteristic (CMC) curves (reporting Rank-1 and Rank-5 accuracies) and mean Average Precision (mAP) as our primary metrics. All evaluations are executed under the standard single-query setting without utilizing re-ranking techniques.

4.2. Implementation Details

Our framework is built upon the CLIP-ViT-B/16 backbone, inheriting the base architectural configurations from ProFD [1]. The visual encoder processes input images resized to 256×128 . Text prompts for $K = 8$ human body parts are initialized according to standard anatomical semantics.

The entire network is trained end-to-end for 60 epochs using the Adam optimizer. The base learning rate is set to 5×10^{-6} for the pre-trained CLIP visual encoder and 3.5×10^{-4} for the randomly initialized modules, modulated by a warm-up cosine annealing scheduler. In our joint optimization objective, we systematically set the spatial distillation weight $\lambda_{\text{distill}} = 0.02$, the feature self-reconstruction weight $\lambda_{\text{recon}} = 0.1$, and the KL divergence penalty $\lambda_{\text{KL}} = 0.01$. For the inference stage, the visibility gating threshold is configured as $\tau_v = 0.2$, while the safe feature substitution threshold is strictly bounded at $S_{\text{IAI}} < 0.1$.

4.3. Comparison with State-of-the-Art Methods

To validate the effectiveness of our design, we benchmark the proposed framework against representative state-of-the-art (SOTA) approaches. The competitors include recent Transformer-based architectures, explicit feature recovery models, and Vision-Language (V-L) prompt frameworks.

Table 1

Comparison with State-of-the-Art (SOTA) Methods on Occluded-ReID/Occluded-Duke Datasets: * denotes models with Transformer backbone, † indicates adding sie_camera, and ‡ indicates encoders with a small-step sliding-window setting (Stride ≤ 13). Bold values = best performance, underlined values = second-best.

Methods	Occluded-ReID		Occluded-Duke	
	R1	mAP	R1	mAP
PVPM [2]	66.8	59.5	-	-
CAAO [24]	65.2	61.1	67.8	55.8
CAAO*	87.1	83.4	68.5	59.5
PFD*† [20]	79.8	81.3	67.7	60.1
TransReID* [5]	70.2	67.3	64.2	55.7
TransReID*‡	-	-	66.4	59.2
PAT* [7]	81.6	72.1	64.5	53.6
FED* [21]	86.3	79.3	68.1	56.4
DPM* [16]	80.2	75.2	66.7	57.2
SGFNet*‡ [9]	93.0	90.3	76.5	67.2
THCB-Net* [18]	87.3	84.5	72.3	62.6
RGANet* [4]	86.4	80.0	71.6	62.4
ProFD* [1]	91.0	88.3	68.9	61.0
SPT*‡ [17]	87.8	81.1	<u>74.7</u>	63.0
MLRAT*‡ [8]	88.9	83.6	73.3	63.1
Ours	<u>93.3</u>	<u>91.3</u>	71.9	63.5
Ours†	-	-	72.2	63.7
Ours‡	93.4	91.9	73.1	<u>64.9</u>

4.3.1. Results on Occluded Benchmarks

As reported in Table 1, we evaluate our method on the Occluded-ReID and Occluded-Duke datasets. On the highly challenging Occluded-Duke dataset, our proposed framework achieves a highly competitive performance of 71.9% Rank-1 and 63.5% mAP under the standard setting, and further reaches 93.3% Rank-1 and 91.3% mAP on Occluded-ReID.

Most importantly, compared to our direct vision-language baseline, ProFD [1] (68.9% Rank-1 and 61.0% mAP), our method yields substantial improvements of +3.0% in Rank-1 and +2.5% in mAP under the exact same standard configuration. While some recent architectures such as SGFNet*‡ [9], SPT*‡ [17], and MLRAT*‡ [8] report higher Rank-1 scores, they heavily rely on small-step overlapping sliding windows (‡, stride ≤ 13) or camera embeddings (†), which substantially increase patch sequences and computational overhead. Notably, under the fair standard non-overlapping setting, our framework achieves the highest mAP (63.5%) among all competitors, even surpassing THCB-Net* [18] (62.6% mAP) which utilizes a larger 384×128 input resolution. Furthermore, when equipped with the same camera embedding and sliding-window settings (**Ours**†‡), our model reaches 73.1% Rank-1 and 64.9% mAP, outperforming both SPT*‡ (63.0% mAP) and MLRAT*‡ (63.1% mAP) by a clear margin. By explicitly mitigating generative uncertainty through our VAE-based feature compensator, our method provides a lightweight, semantically interpretable, and mathematically elegant solution.

Table 2

Comparison with State-of-the-Art (SOTA) Methods on Market1501/DukeMTMCreID

Methods	Market1501		DukeMTMC	
	R1	mAP	R1	mAP
CAAO [24]	95.1	87.3	88.9	77.5
CAAO*	95.3	88.0	89.8	80.9
PFD* [†] [20]	95.5	89.6	90.6	82.2
TransReID* [†] [5]	95.0	88.2	89.6	80.6
PAT* [7]	95.4	88.0	88.8	78.2
FED* [21]	95.0	86.3	89.4	78.0
SGFNet* [‡] [9]	<u>96.2</u>	<u>91.1</u>	92.8	84.1
THCB-Net* [18]	<u>96.2</u>	90.6	91.7	83.5
RGANet* [4]	95.5	89.8	-	-
ProFD* [1]	95.1	90.0	91.7	83.2
SPT* [‡] [17]	95.5	89.4	91.1	82.4
MLRAT* ^{†‡} [8]	95.8	89.9	-	-
Ours	95.7	90.3	91.8	84.1
Ours[†]	95.8	90.5	92.8	<u>85.5</u>
Ours^{†‡}	96.4	91.3	<u>92.7</u>	86

4.3.2. Results on Holistic Benchmarks

A robust occlusion-handling mechanism should safely revert to standard processing when observing uncorrupted inputs. To verify that our design does not degrade discriminative capabilities in standard scenarios, we evaluate the framework on Market-1501 and DukeMTMC-ReID.

As shown in Table 2, our method achieves highly competitive performance on both holistic benchmarks, reaching 95.7% Rank-1 and 90.3% mAP on Market-1501, and 91.8% Rank-1 and 84.1% mAP on DukeMTMC-ReID under the standard setting. Furthermore, equipped with camera embeddings and sliding-window settings (**Ours^{†‡}**), our model achieves state-of-the-art results of 96.4% Rank-1 (91.3% mAP) on Market-1501 and 92.7% Rank-1 (86.0% mAP) on DukeMTMC-ReID. This confirms that our Uncertainty-guided Dynamic Retrieval (UDR) assigns high visibility confidence to unoccluded regions, preserving and further enhancing global representation quality.

4.4. Ablation Studies

(Note: To cleanly verify the architectural contributions without the interference of advanced sliding-window or camera-embedding tricks, all ablation studies are conducted using the standard baseline configuration.)

4.4.1. Comprehensive Component-wise Ablation

To rigorously validate the individual and synergistic contributions of our proposed modules, we conduct a comprehensive ablation study, as summarized in Table 3.

Table 3: Comprehensive component-wise ablation study of the proposed framework on Occluded-Duke. Best results are highlighted in **bold**.

Method	$\mathcal{L}_{distill}$	$\mathcal{L}_{recon}$	$\mathcal{L}_{KL}$	mAP (%)	Rank-1 (%)	Rank-5 (%)
Baseline				60.99	68.87	82.08
+ TDPP (Distillation)	✓			61.85	69.75	82.65
+ PMFC (Deterministic AE)		✓		62.22	70.09	83.17
+ PMFC (Variational VAE)		✓	✓	62.98	71.35	83.74
+ TDPP & PMFC (AE)	✓	✓		62.57	70.51	83.35
+ Full Framework (Ours)	✓	✓	✓	63.46	71.87	83.91

Individual Effectiveness: When applied independently, the visibility-aware spatial distillation ($\mathcal{L}_{distill}$) in TDPP provides a modest but reliable improvement over the baseline (+0.88% Rank-1 and +0.86% mAP), confirming its role as a robust semantic localizer. Meanwhile, the standalone PMFC modules contribute more substantially. The deterministic AE version yields 70.09% Rank-1, while the full variational formulation (VAE) leverages generative uncertainty to reach an impressive 71.35% Rank-1, serving as the core engine for occlusion handling.

Synergistic Integration: Finally, we evaluate the joint framework. Combining TDPP with the deterministic AE version of PMFC achieves 70.51% Rank-1. More importantly, when the full variational framework is assembled (TDPP + VAE), the model successfully reaches its peak at 71.87% Rank-1 and 63.46% mAP. Notably, we observe a highly consistent and significant boost across the core metrics: both Rank-1 and mAP surge by approximately 0.5% (+0.52% and +0.48%, respectively) over the strong VAE-only configuration. This demonstrates that the synergy of explicit semantic addressing (TDPP) and uncertainty-guided substitution (PMFC) robustly refines the global ranking quality. Concurrently, the saturation in Rank-5 (+0.17% gain) reflects the inherent difficulty of the remaining severely corrupted tail samples.

4.4.2. Hyperparameter Sensitivity Analyses

To isolate the parameter behavior of each module without cross-module interference, we independently dissect the sensitivity of the three critical optimization objectives on their respective standalone branches (matching the individual settings in Table 3).

Table 4: Sensitivity analysis of the reconstruction loss weight λ_{recon} on Occluded-Duke. Evaluated in the deterministic reconstruction setting (λ_{recon} varied with $\lambda_{KL} = 0$).

Weight (λ_{recon})	mAP (%)	Rank-1 (%)	Rank-5 (%)
0 (Baseline)	60.99	68.87	82.08
1×10^{-4}	61.58	69.45	82.60
1×10^{-3}	61.85	69.73	82.85
1×10^{-2}	62.06	69.91	83.02
1×10^{-1}	62.22	70.09	83.17
2×10^{-1}	61.95	69.64	82.76

Table 5: **Sensitivity analysis of the spatial distillation weight $\lambda_{distill}$ for TDPP on Occluded-Duke.** The distillation objective provides a steady foundational enhancement, peaking at $\lambda_{distill} = 0.02$.

Weight ($\lambda_{distill}$)	mAP (%)	Rank-1 (%)	Rank-5 (%)
0 (Baseline)	60.99	68.87	82.08
1.0×10^{-4}	61.25	69.15	82.20
1.0×10^{-3}	61.42	69.35	82.35
1.5×10^{-2}	61.70	69.62	82.52
2.0×10^{-2}	61.85	69.75	82.65
1.0×10^{-1}	61.50	69.30	82.35

Table 6: **Sensitivity analysis of the KL divergence weight λ_{KL} on Occluded-Duke (under active feature compensation).**

Weight (λ_{KL})	mAP (%)	Rank-1 (%)	Rank-5 (%)
0 (Recon only)	62.22	70.09	83.17
1×10^{-4}	62.48	70.55	83.35
1×10^{-3}	62.74	70.92	83.56
1×10^{-2}	62.98	71.35	83.74
1×10^{-1}	61.88	69.64	82.75

- **Reconstruction Weight (λ_{recon}):** As shown in Table 4, configuring $\lambda_{recon} = 0.1$ delivers the optimal trade-off. When λ_{recon} is excessively small ($< 10^{-3}$), the geometric constraint is insufficient to guide missing token prediction. Conversely, when $\lambda_{recon} > 0.1$, the reconstruction objective overly dominates the optimization landscape, which slightly interferes with the discriminative metric learning.
- **Distillation Weight ($\lambda_{distill}$):** In Table 5, the performance peaks at $\lambda_{distill} = 0.02$. While this spatial regularization establishes a geometrically accurate representation anchor, applying an overly strict penalty (≥ 0.1) forces the textual proxies to rigidly fit the pseudo-parsed masks, thereby reducing their adaptability to varying poses.
- **KL Divergence Penalty (λ_{KL}):** The variational constraint on the latent space is analyzed in Table 6. Because the UDR mechanism heavily relies on a well-behaved posterior variance, scaling λ_{KL} from the deterministic baseline to 0.01 unlocks the true potential of uncertainty-guided retrieval (71.35% Rank-1). However, excessively large penalties (≥ 0.1) trigger posterior over-smoothing, causing the generated features to lose their identity-specific discriminability.

4.5. Qualitative Visualizations

Feature Space Distribution via t-SNE (Figure 2): We visualize the global representation manifolds optimized by the ProFD baseline and our framework. It is evident that our model forms tighter intra-identity clusters and maintains clearer inter-identity margins under occlusion. This confirms that semantic-guided evidence aggregation, combined with robust feature compensation, effectively suppresses occlusion-induced feature drift.

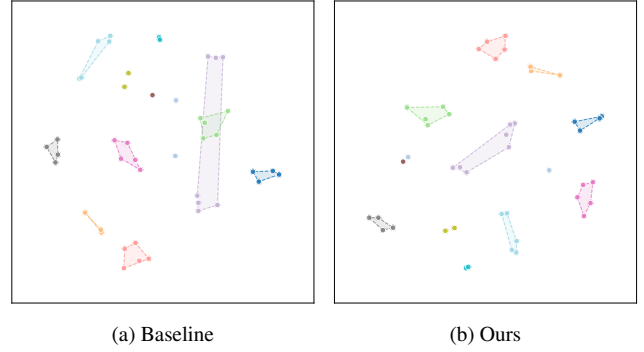


Figure 2: **t-SNE visualization of the learned feature spaces.** Different colors denote different pedestrian identities. Compared with the ProFD baseline, the proposed framework produces more compact intra-identity distributions and clearer inter-identity separation under occlusion. The improvement suggests that semantic-guided part aggregation, masked feature compensation, and uncertainty-guided dynamic retrieval jointly contribute to a more discriminative and structurally consistent representation space.

Hard-Sample Retrieval Rankings (Figure 3): We compare the top-ranked gallery returns for severely occluded queries. In cases where baseline matches completely fail due to heavy foreground obstacles (e.g., pillars or passing pedestrians), our framework successfully identifies the correct identity. This highlights the practical advantage of using generative uncertainty to safely replace corrupted tokens without injecting hallucinated noise into the distance metric.



Figure 3: Visualization of retrieval results on challenging occluded queries. The leftmost column shows severely occluded query images, followed by the top-ranked gallery results obtained by the ProFD baseline and our method. Green and red bounding boxes indicate correct and incorrect matches, respectively. Compared with the baseline, our method maintains more reliable identity matching when discriminative body regions are partially or heavily occluded, benefiting from semantic-guided evidence aggregation and feature compensation for incomplete local observations.

Complexity (Figure 4): The proposed TDPP and PMFC achieve substantial performance gains with negligible computational overhead. Specifically, our modules add only 0.33 GFLOPs in total, preserving the inference efficiency of the baseline model.

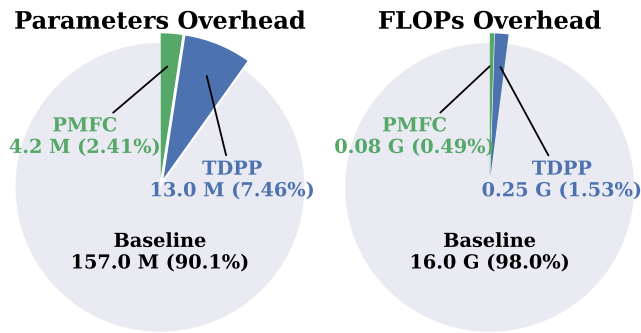


Figure 4: **Computational complexity overhead.** Our proposed modules (TDPP and PMFC) incur minimal additional parameters (M) and FLOPs (G) compared to the ProFD baseline, ensuring highly efficient occlusion recovery with nearly zero extra computational cost.

5. Conclusion

In this paper, we address the severe information loss and metric corruption issues in occluded person re-identification by rethinking feature recovery as a variational, semantic-guided process. We present a unified framework that overcomes the perceptual limitations of standard discriminative models through a "probe-and-compensate" paradigm. By introducing Text-Driven Proxy Pooling (TDPP), we establish robust semantic anchors to precisely aggregate visible part evidence. To rescue the missing topologies, our Prompt-Guided Masked Feature Compensation (PMFC) steps beyond deterministic guessing by formulating the recovery as a conditional variational problem. Crucially, the resulting generative posterior variance naturally unlocks the Uncertainty-guided Dynamic Retrieval (UDR) mechanism. This design effectively evaluates the reliability of compensated features, acting as a smart gradient gate that seamlessly supplements critical topological details while rigorously shielding the identity metric space from hallucinated noise. Extensive evaluations validate that our method not only sets new state-of-the-art records on challenging occluded benchmarks but also offers profound insights into harmonizing generative uncertainty with discriminative metric learning. We believe this uncertainty-aware retrieval perspective will inspire future robust vision-language applications in complex real-world environments.

References

- [1] Cui, C., Huang, S., Song, W., Ding, P., Zhang, M., Wang, D., 2024. Profd: Prompt-guided feature disentangling for occluded person re-identification, in: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 1583–1592.
- [2] Gao, S., Wang, J., Lu, H., Liu, Z., 2020. Pose-guided visible part matching for occluded person reid, in: Proceedings of the IEEE/CVF

- conference on computer vision and pattern recognition, pp. 11744–11752.
- [3] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), IEEE. pp. 15979–15988.
- [4] He, S., Chen, W., Wang, K., Luo, H., Wang, F., Jiang, W., Ding, H., 2023. Region generation and assessment network for occluded person re-identification. IEEE transactions on information forensics and security 19, 120–132.
- [5] He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W., 2021. Transreid: Transformer-based object re-identification, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 15013–15022.
- [6] Li, S., Sun, L., Li, Q., 2023. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels, in: Proceedings of the AAAI conference on artificial intelligence, pp. 1405–1413.
- [7] Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F., 2021. Diverse part discovery: Occluded person re-identification with part-aware transformer, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 2897–2906.
- [8] Lin, G., Bao, Z., Huang, Z., Li, Z., Zheng, W.S., Chen, Y., 2024. A multi-level relation-aware transformer model for occluded person re-identification. Neural Networks 177, 106382.
- [9] Lin, G., Yang, S., Zheng, W.S., Li, Z., Huang, Z., 2025a. A semantically guided and focused network for occluded person re-identification. IEEE Transactions on Information Forensics and Security .
- [10] Lin, G., Zheng, W.S., Li, Z., Lu, Y., Ma, X., Huang, Z., 2025b. Vision-language models for person re-identification: a survey and outlook. Information Fusion , 104095.
- [11] Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y., 2019. Pose-guided feature alignment for occluded person re-identification, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE. pp. 542–551.
- [12] Ning, E., Wang, C., Zhang, H., Ning, X., Tiwari, P., 2024. Occluded person re-identification with deep learning: a survey and perspectives. Expert systems with applications 239, 122419.
- [13] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR. pp. 8748–8763.
- [14] Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking, in: European conference on computer vision, Springer. pp. 17–35.
- [15] Sun, S., Lin, J., Weng, J., Liu, Z., Li, S., Luo, Z., 2026. Cope: Consistent occlusion and prompt enhancement network for occluded person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11293–11302.
- [16] Tan, L., Dai, P., Ji, R., Wu, Y., 2022. Dynamic prototype mask for occluded person re-identification, in: Proceedings of the 30th ACM international conference on multimedia, pp. 531–540.
- [17] Tan, L., Xia, J., Liu, W., Dai, P., Wu, Y., Cao, L., 2024. Occluded person re-identification via saliency-guided patch transfer, in: Proceedings of the AAAI conference on artificial intelligence, pp. 5070–5078.
- [18] Wang, C., Gao, X., Wu, M., Lam, S.K., He, S., Tiwari, P., 2025. Looking clearer with text: A hierarchical context blending network for occluded person re-identification. IEEE Transactions on Information Forensics and Security 20, 4296–4307.
- [19] Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J., 2020. High-order information matters: Learning relation and topology for occluded person re-identification, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 6448–6457.

- [20] Wang, T., Liu, H., Song, P., Guo, T., Shi, W., 2022a. Pose-guided feature disentangling for occluded person re-identification based on transformer, in: Proceedings of the AAAI conference on artificial intelligence, pp. 2540–2549.
- [21] Wang, Z., Zhu, F., Tang, S., Zhao, R., He, L., Song, J., 2022b. Feature erasing and diffusion network for occluded person re-identification, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 4744–4753.
- [22] Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H., 2022. Simmim: A simple framework for masked image modeling, in: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), IEEE. pp. 9643–9653.
- [23] Xu, B., He, L., Liang, J., Sun, Z., 2022. Learning feature recovery transformer for occluded person re-identification. *IEEE Transactions on Image Processing* 31, 4651–4662.
- [24] Zhao, C., Qu, Z., Jiang, X., Tu, Y., Bai, X., 2023. Content-adaptive auto-occlusion network for occluded person re-identification. *IEEE Transactions on Image Processing* 32, 4223–4236.
- [25] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q., 2015. Scalable person re-identification: A benchmark, in: Proceedings of the IEEE international conference on computer vision, pp. 1116–1124.
- [26] Zhuo, J., Chen, Z., Lai, J., Wang, G., 2018. Occluded person re-identification, in: 2018 IEEE international conference on multimedia and expo (ICME), IEEE. pp. 1–6.