Reliability-Aware 3D Geometric Injection for Universal Person Re-identification

Bohan Su[1*], Jiashuo Wang[1*], Fangyi Liu[1†], and Mang Ye[1†]

[1] National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, China
{bohansu,jiashuowang,fangyiliu,yemang}@whu.edu.cn

Abstract. Universal person re-identification (ReID) aims to retrieve pedestrian identities across diverse real-world scenarios, including severe occlusions, clothing changes, and cross-modality shifts, within a unified model. However, existing 2D representations fundamentally struggle with spatial ambiguities due to a lack of depth and topological awareness, while naively introducing monocular 3D priors often causes severe negative transfer due to geometric estimation noise under extreme visual degradation. To safely harness the clothing-invariant and canonical structural properties of 3D geometry, we propose UniGeo, a Universal Monocular 3D-Enhanced ReID framework driven by a Consistency-Aware Reliability Gate and Dual-Stream Residual Fusion. Specifically, the processing of 3D information is strategically decoupled into geometric extraction and dynamic utilization. To provide pure structural compensation, we project monocular 3D parameters into kinematic joint representations, explicitly capturing instance-level geometric topology to resolve appearance-based ambiguities. To robustly incorporate these cues without perturbing the reliable 2D feature space, we isolate the 3D prior as a late-stage structural residual; modulated by the consistency-aware gate, this mechanism adaptively filters geometric noise and enables controlled fallback to the pure 2D baseline. Extensive experiments show that our method improves challenging, structure-sensitive scenarios while preserving competitive performance on clean domains. Code is available at https://github.com/BohanSu/UniGeo.

Keywords: Universal Person ReID · 3D · Reliability Gate

1 Introduction

Person re-identification (ReID) aims to retrieve the same individual across non-overlapping camera views, serving as a fundamental capability for large-scale video surveillance systems and long-term identity tracking applications [87, 98]. While recent deep learning baselines have achieved remarkable progress in closed-set scenarios with controlled conditions [21, 53], real-world deployments demand robust identification under severe and unpredictable appearance corruptions.

* Equal contribution. † Co-corresponding authors.

These corruptions span a wide spectrum of challenges, including partial occlusions caused by obstacles or crowds [54, 96], dramatic clothing changes across sessions [20,85], and cross-sensor variations in illumination, resolution, and even imaging modality (*e.g.* visible-to-infrared transitions) [79, 87].

This practical imperative motivates the paradigm of *Universal ReID* [3,100]: instead of maintaining separate specialist models for each scenario, the goal is one model that generalizes across heterogeneous domains. Maintaining parallel expert models is expensive, architecturally brittle, and difficult to scale in practice. These domains include standard holistic benchmarks such as Market-1501 [97] and MSMT17 [78], clothing-change scenarios such as PRCC [85], severe occlusion benchmarks such as Occluded-Duke [54], and cross-modality datasets such as SYSU-MM01 [79]. The promise of Universal ReID is to reduce task-specific silos and support seamless deployment across varied real-world conditions. Yet universality remains difficult because these domains differ in both low-level visual statistics (viewpoint, style, resolution) and semantic structures (pose articulation, body topology, and modality characteristics) [10, 75].

Recent 2D advances include transformer-based representation learning [21, 104], vision-language transfer [37], and prompt-driven universal adaptation [100]. Together, these methods provide increasingly strong visual foundations for cross-domain generalization. Yet a critical limitation persists: their identity evidence still relies mainly on 2D appearance cues from textures and color patterns [18]. When such cues fail under heavy occlusion [84], extreme lighting variation [79], or clothing change [23], these representations suffer spatial ambiguities and semantic collapse. Consequently, as illustrated in Fig. 1(a), current universal pipelines still face a structural bottleneck when appearance information is heavily corrupted or uninformative [4].

Introducing 3D geometry is therefore a principled way to overcome this bottleneck [4, 103]. Recent advances in monocular 3D human reconstruction, especially transformer-based models such as 4DHumans [19], have made geometry extraction practical for ReID. We adopt 4DHumans as our default SMPL extractor following recent practice, while keeping the framework model-agnostic: any SMPL-based estimator can provide the upstream geometry. The core challenge is not choosing a 3D estimator, but using its outputs safely. Monocular 3D reconstruction remains inherently ill-posed [93], and even state-of-the-art estimators become unreliable under severe visual degradations (blur, truncation, heavy occlusion) or domain shifts (*e.g.* RGB-to-infrared transitions in cross-modality scenarios). Naive static fusion that unconditionally injects such noisy estimates can propagate geometric errors into the feature space, ultimately degrading rather than improving ReID discrimination performance [66].

To overcome these limitations, as depicted in Fig. 1(b), our main contribution is to treat 3D geometric cues as *conditional* structural evidence rather than a uniformly trusted auxiliary input [32]. We instantiate this paradigm through a two-part architecture: (1) a kinematic-aware geometry extraction branch that projects off-the-shelf SMPL parameters [52] into structured pose representations $\mathbf{f}_{pose}$ capturing body topology, and (2) a reliability-aware gating mechanism that

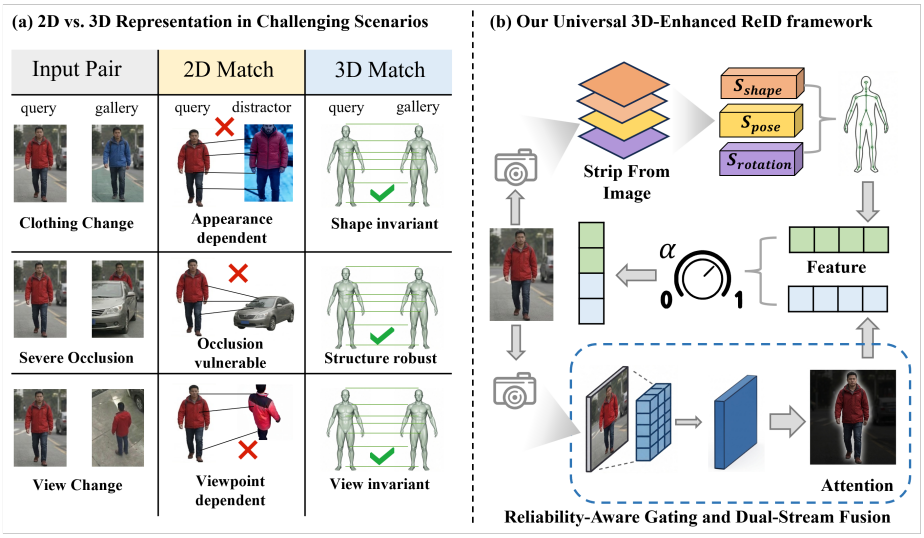

Fig. 1: (a) Standard 2D pipelines are vulnerable to severe appearance corruptions (*e.g.* heavy occlusion, clothing changes). (b) Our 3D-aided pipeline leverages kinematic-aware geometry and reliability-aware gating to conditionally fuse 3D structural cues with 2D features, ensuring robust identity representation.

dynamically estimates geometric reliability by evaluating cross-modal consistency [60]. This gating mechanism provides robustness even in demanding cases, such as cross-modality inputs where RGB-trained 3D estimators can become unreliable. It achieves this by detecting visual-geometric discrepancies and safely suppressing unreliable structural signals, thereby preventing negative transfer while preserving strong 2D baselines.

In summary, our key contributions are threefold:

- We reformulate monocular 3D geometry as *conditional structural evidence* for Universal ReID, rather than as a replacement for 2D appearance. This perspective is designed for cases where appearance cues become ambiguous under occlusion, clothing change, or modality shift.
- We propose UniGeo, which extracts a kinematic-aware structural representation from off-the-shelf SMPL estimates [19, 52] and injects it through a consistency-aware gate. The model activates geometry when it is helpful and suppresses unreliable 3D when visual-geometric mismatch indicates high risk of negative transfer.
- UniGeo delivers clear improvements where structural cues matter most, including +3.0% Rank-1 on PRCC, +0.9% on Occluded-Duke, and +1.2% on SYSU-MM01, while remaining competitive on standard benchmarks such as Market-1501, MSMT17, and CUHK03. These results confirm that reliability-aware geometric injection strengthens robustness under domain corruption without sacrificing standard-domain performance.

2 Related Work

From Specific to Universal ReID. Traditional ReID architectures yield strong performance under closed-world assumptions [21, 37, 53, 87]. To address real-world domain shifts, existing works mainly target isolated challenges: part-aware and pose-guided alignments for severe occlusion [17, 42, 54, 67, 77, 84, 103]; shape-aligned and gait-assisted disentanglement for clothing changes [6, 20, 23, 29, 47, 48, 59, 85]; and modality compensation for discrepancies [15, 63, 79, 87].

While these specialized models excel individually, deploying parallel expert models for unpredictable visual corruptions is impractical. Recent universal frameworks, such as VersReID [100] and UNIReID [3], unify diverse domains by routing features through scene-specific prompts or task-aware learning, while recent cloth-generalized studies further emphasize robustness under severe appearance shifts [47, 48]. However, these approaches are still primarily optimized around appearance-driven domain adaptation. As a complementary perspective, we advance the universal paradigm by exploring how 3D auxiliary information can resolve microscopic instance-level structural distortions when texture cues fail. This perspective is particularly relevant when body layout and pose geometry remain semantically informative even after clothing textures or modality-specific color statistics become unreliable. By dynamically injecting monocular 3D priors through a reliability-aware gating mechanism, our dual-stream framework provides robust structural compensation across deployments.

Monocular 3D Human Reconstruction. Recovering 3D human parametric models (*e.g.* SMPL [1, 52]) from monocular images provides a solid foundation for extracting structural priors robust to 2D pixel corruption. The field has rapidly evolved from early end-to-end regression frameworks [31, 58] to methods exploiting pixel-aligned supervision [34, 55, 92] and spatial robustness mechanisms [27, 33, 43, 105] for handling complex articulations. Recently, Transformer-based architectures [9, 44–46] have predominated, culminating in state-of-the-art foundation models like 4DHumans [19].

Despite this progress, in-the-wild monocular 3D recovery remains fundamentally ill-posed. As highlighted by probabilistic models [35] and recent surveys [93], 2D projection introduces depth ambiguity, which is further amplified by severe occlusions and complex scenes. As a result, geometric predictions often become unstable near object boundaries and occluded regions. This intrinsic uncertainty introduces estimation noise, so 3D outputs must be integrated carefully when serving downstream recognition tasks.

3D-Assisted Person Re-identification. Historically, 3D information was introduced for video sequence aggregation [80], later expanding to novel modalities like WiFi [64]. In RGB/IR domains, early works leverage 3D models for data augmentation [68, 74, 94, 101]. To tackle clothing changes, while some baselines mine invariants from RGB modalities only [20], 3D-assisted methods leverage meshes, multi-view transformations [89], and skeleton dynamics [30, 61, 62] for

shape-aware alignment [4, 54, 63, 66]. Recent advances further establish 2D-3D correspondences [76] or learn clothing-invariant 3D representations [49].

While robust against specific texture variations, these pioneering works assume reliable geometric estimation. However, under unpredictable visual corruptions in Universal ReID, off-the-shelf estimators can produce distorted structures, and unconditionally injecting them can induce negative transfer. Advancing a noise-tolerant philosophy, our approach differs from static or dynamic skeleton encodings [30,61,62] and deterministic 3D fusions [49]. Instead of rigid 2D-3D correspondences [76] or pure RGB mining [20], our Reliability-Aware Geometric Injection utilizes kinematic decoupling and a consistency-aware gating mechanism to dynamically modulate 3D fusion. This selectively harnesses geometric invariants when reliable, while limiting estimation noise during extreme spatial-texture conflicts and improving robustness in cross-modality settings [64].

3 Method

The overall pipeline of our proposed UniGeo framework is illustrated in Fig. 2. With the core objective of extracting robust representations in unpredictable environments, our architecture strategically decouples feature learning into two synergistic streams, safely integrating 3D geometry as structural compensation.

In this section, we first introduce the Scene-Aware Visual Stream (Sec. 3.1), which provides a 2D baseline for resolving macroscopic domain shifts. We then detail the 3D Auxiliary Structural Stream (Sec. 3.2), which extracts kinematic-aware 3D topological features to mitigate instance-level corruptions such as severe occlusions. Next, given the inherent instability of monocular 3D estimators in the wild [33, 93], we present a Reliability-Aware Geometric Gate and a Dual-Stream Fusion strategy (Sec. 3.3) to prevent geometric noise propagation. Finally, we describe the end-to-end optimization scheme and cached-geometry deployment and inference strategy (Sec. 3.4).

3.1 Scene-Aware Visual Representation

In real-world scenarios characterized by macroscopic data distribution discrepancies (*e.g.* illumination shifts), conventional 2D backbones often struggle to disentangle identity features from environmental noise. Inspired by prompt tuning, we establish a Scene-Aware Visual Stream based on a Vision Transformer (ViT) [14,100] to explicitly incorporate environmental contexts, thereby extracting domain-resilient appearance textures.

Formally, an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ is partitioned and linearly projected into patch embeddings $\mathbf{X}_{patch} \in \mathbb{R}^{N\times D}$, where D is the embedding dimension. To achieve scene-awareness, we introduce learnable scene-specific prompts $\mathbf{P}_{scene} \in \mathbb{R}^{K\times D}$ (K denotes the prompt length). Following the protocol in VersReID [100], these prompts are conditioned on dataset-level scene labels rather than specific camera IDs or raw environmental tags; specifically, each dataset is

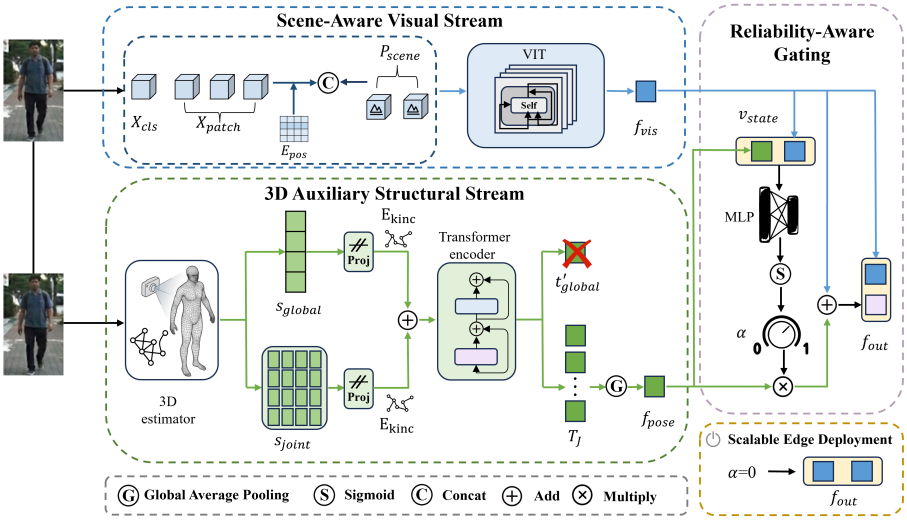

Fig. 2: Overview of our UniGeo framework. A Scene-Aware Visual Stream (Sec. 3.1) extracts 2D textures $\mathbf{f}_{vis}$, while an Auxiliary Structural Stream (Sec. 3.2) models decoupled 3D geometric features $\mathbf{f}_{pose}$. Guided by cross-modal discrepancy, a Reliability-Aware Gate predicts scalar α to modulate the dual-stream residual fusion (Sec. 3.3) under end-to-end optimization (Sec. 3.4).

assigned a scene label denoting its primary ReID challenge (*e.g.* general, low-resolution, occlusion). During multi-dataset training, these labels guide scene-prompt selection to encode domain knowledge. Prepended with a class token $\mathbf{x}_{cls} \in \mathbb{R}^{1\times D}$, the patch tokens are first supplemented with spatial position embeddings $\mathbf{E}_{pos} \in \mathbb{R}^{(1+N)\times D}$. The scene prompts are then appended to the sequence end:

$$\mathbf{Z}_0 = \big[[\mathbf{x}_{cls}; \mathbf{X}_{patch}] + \mathbf{E}_{pos};\ \mathbf{P}_{scene}\big]. \tag{1}$$

After processing through L Transformer layers, the state of the class token is extracted as the global visual representation $\mathbf{f}_{vis} \in \mathbb{R}^D$.

While effective against macroscopic domain shifts, this 2D representation inherently relies on local textures and appearance cues [18]. Consequently, under severe instance-level structural corruptions (*e.g.* heavy occlusions), it inevitably suffers from spatial ambiguities due to the lack of topological awareness, which naturally motivates our 3D structural branch.

3.2 Geometric Feature Extraction

As analyzed above, 2D visual representations inevitably collapse under severe instance-level corruptions, demanding robust structural compensation. To this end, we introduce a lightweight 3D structural branch utilizing an off-the-shelf

monocular 3D estimator, 4DHumans [19], to extract the SMPL [52] parameter vector $\mathbf{s} \in \mathbb{R}^{82}$. Mathematically, this parameter vector encapsulates highly heterogeneous physical quantities, including global spatial orientation and local joint articulations. Treating it merely as a homogeneous black-box vector processed by a naive multi-layer perceptron (MLP) would overlook the explicit physical semantics and the kinematic dependencies among body parts. Motivated by this, we design a kinematic-aware Pose Encoder (E_{pose}) to explicitly decouple these components and model their structural topology.

Specifically, to preserve physical semantics, we decouple the 82-dimensional SMPL vector into two components: global parameters $\mathbf{s}_{global} \in \mathbb{R}^{13}$, which encode 3D global rotation and 10-D shape proportions, and local joint parameters $\mathbf{s}_{joint} \in \mathbb{R}^{23\times 3}$, which represent the relative 3D rotations of 23 body joints.

Instead of mixing these disparate spaces prematurely, two independent projection pathways (linear layers followed by GELU) map them into the latent dimension D. This yields a global token $\mathbf{t}_{global} \in \mathbb{R}^{1\times D}$ and a sequence of local joint tokens $\mathbf{T}_{local} \in \mathbb{R}^{23\times D}$. To capture kinematic topology, we concatenate these tokens, add positional embeddings $\mathbf{E}_{kine} \in \mathbb{R}^{(1+23)\times D}$, and process the sequence with a lightweight Transformer encoder:

$$[\mathbf{t}'_{global}; \mathbf{T}'_{local}] = \text{Transformer}\big([\mathbf{t}_{global}; \mathbf{T}_{local}] + \mathbf{E}_{kine}\big). \tag{2}$$

Global camera shifts and absolute spatial coordinates often introduce severe view-dependent interference in wild environments. To reduce this variance and focus on articulated human posture, we adopt a two-stage design: during Transformer processing, the global parameters $\mathbf{s}_{global}$ (containing identity-relevant shape factors such as limb proportions and body structure) are fed as input to enrich the joint tokens through cross-attention; however, for the final output, we discard the updated global token $\mathbf{t}'_{global}$ and retain only the 23 refined joint tokens, denoted as $\mathbf{T}_J \in \mathbb{R}^{J\times D}$ (where $J = 23$), to improve view invariance. This configuration—using both $\mathbf{s}_{global}$ and $\mathbf{s}_{joint}$ as Transformer input while pooling $\mathbf{f}_{pose}$ exclusively from joint tokens—is validated as the optimal topology in our ablation study (Tab. 5). Subsequently, a global average pooling operation is applied across the joint token sequence to distill a compact geometric feature:

$$\mathbf{f}_{pose} = \frac{1}{J}\sum_{j=1}^{J} \mathbf{t}_{J,j}, \tag{3}$$

where $\mathbf{t}_{J,j} \in \mathbb{R}^D$ denotes the j-th joint token within the sequence $\mathbf{T}_J$. Independent of fragile 2D textures, $\mathbf{f}_{pose}$ provides an explicit structural anchor, laying a solid foundation for the subsequent reliability-aware fusion.

3.3 Reliability-Aware Gating and Dual-Stream Fusion

While the aforementioned 3D prior offers valuable topological compensation, monocular 3D reconstruction is fundamentally an ill-posed problem. Under severe visual degradations (*e.g.* heavy truncations), the estimator inevitably produces erroneous geometries. Blindly fusing such geometric noise would corrupt

the identity representation, leading to severe negative transfer. To systematically prevent this, we hypothesize that the reliability of the 3D estimation is highly correlated with its spatial consistency with the 2D visual semantics. Consequently, we propose a Reliability-Aware Geometric Gate coupled with a Dual-Stream Residual Fusion strategy to achieve risk-aware structural integration.

Consistency-Aware Reliability Gate. To explicitly evaluate the cross-modal discrepancy, we concatenate the global visual representation and the geometric feature to form a joint state vector $\mathbf{v}_{state} = [\mathbf{f}_{vis}; \mathbf{f}_{pose}] \in \mathbb{R}^{2D}$. Rather than relying on heuristic distance metrics, $\mathbf{v}_{state}$ is fed into a lightweight multi-layer perceptron (MLP) to implicitly model the geometry-visual alignment and predict a consistency-aware reliability coefficient α:

$$\alpha = \sigma\left(\mathrm{MLP}(\mathbf{v}_{state})\right) \in (0, 1), \tag{4}$$

where $\sigma(\cdot)$ denotes the Sigmoid function. To maintain computational efficiency, the MLP employs a bottleneck architecture, projecting an input of dimension $2D$ into a compact hidden space of dimension $D/4$ before mapping it to the final scalar. Driven by end-to-end optimization, this gate acts as an adaptive reliability filter, suppressing α when unresolvable divergence occurs between the 2D texture and the estimated 3D geometry.

Dual-Stream Residual Fusion. To safely integrate the geometric prior without compromising the stability of the 2D baseline, we isolate the 3D structural intervention as a controlled residual. The final ReID representation $\mathbf{f}_{out} \in \mathbb{R}^{2D}$ is formulated as a concatenation-residual hybrid:

$$\mathbf{f}_{out} = \left[\mathbf{f}_{vis};\ \mathbf{f}_{vis} + \alpha \cdot \mathbf{f}_{pose}\right]. \tag{5}$$

This late-stage formulation preserves the stability of the visual stream. The first half of $\mathbf{f}_{out}$ preserves the pure 2D visual feature semantics as an uncorrupted anchor. The second half acts as a dynamically modulated structural residual. When severe visual degradation triggers $\alpha \approx 0$, the structural residual reduces to $\mathbf{f}_{vis}$, completely halting geometric noise propagation. This design lets the dual-stream representation revert to the pure 2D feature space, providing a stable fallback against 3D estimation failures.

3.4 Optimization and Scalable Edge Deployment

To rigorously validate that our performance improvements stem from the proposed geometric gating mechanism rather than auxiliary training tricks, we adhere to a standard ReID optimization paradigm. The dual-stream framework is trained end-to-end, with the overall objective imposed exclusively on the final fused representation $\mathbf{f}_{out}$:

$$\mathcal{L}_{total} = \mathcal{L}_{cls}(\mathbf{p}, y) + \lambda \cdot \mathcal{L}_{tri}(\mathbf{f}_{out}, y), \tag{6}$$

where $\mathbf{p}$ denotes the identity logits from a linear classifier, $\mathcal{L}_{cls}$ is cross-entropy with label smoothing, $\mathcal{L}_{tri}$ is the batch-hard triplet loss, y is the identity label, and λ weights the triplet term.

Cached Geometry and Fallback. Executing monocular 3D estimators on resource-constrained edge devices is often computationally prohibitive. Our late-fusion architecture addresses this through offline-cached geometry extraction and an inference-time fallback mechanism.[1] By manually enforcing $\alpha = 0$, the structural stream is bypassed, degenerating the final representation to $\mathbf{f}_{out} = [\mathbf{f}_{vis}; \mathbf{f}_{vis}]$. Since the cosine similarity between such duplicated vectors is mathematically equivalent to their pure 2D counterparts ($\mathbf{f}_{vis}$), the retrieval behavior remains compatible with the 2D baseline without feature distribution shifts. In practice, this design decouples accuracy-oriented geometric preprocessing from lightweight deployment constraints. When cached geometry is available, the model can exploit structural cues; when it is not, the same retrieval pipeline still operates in a visual-only mode without changing the similarity rule.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on 9 benchmarks grouped into 5 scenarios: **A. Standard Holistic** (Market-1501 [97], MSMT17 [78], and CUHK03 [39] using the 1367/100 labeled split), **B. Clothing-Change** (PRCC [85], Celeb-ReID [24]), **C. Occlusion** (Occluded-Duke [54]), **D. Cross-Modality** (SYSU-MM01 [79]), and **E. UAV & Wild** (UAV-Human [38], AG-ReID.v2 [56]). We report Rank-1 and mAP without re-ranking.

Implementation Details. All models use ViT-Base (384×128) with ImageNet pre-training, trained on a single RTX 3090 GPU throughout all experiments reported in this paper. We employ universal joint training with domain-balanced sampling for 180 epochs using SGD (momentum 0.9, lr=4×10^{-4}, weight decay=1×10^{-4}, cosine annealing), batch size 120 (30 IDs $\times$ 4 images). The loss combines cross-entropy and triplet loss (margin 0.3) with equal weights. SMPL parameters are extracted offline using the pre-trained 4DHumans [19] model (frozen weights, 25 FPS on RTX 3090, $\sim$40ms per image) and cached to disk, eliminating online 3D estimation overhead during training and inference. The 3D stream uses a 2-layer Transformer encoder (hidden dim 768, 8 attention heads). The gating MLP has architecture [1536$\rightarrow$768$\rightarrow$1] with GELU activation. For data augmentation alignment, random horizontal flips are synchronized by mirroring SMPL joint x-coordinates; random cropping (scale 0.75-1.0) and random erasing (p=0.5) apply only to the visual stream, forming an asymmetric augmentation strategy that exposes the gate to spatially misaligned cases during training and improves robustness at inference.

Baseline Configuration. Our pure 2D baseline adopts the VersReID [100] architecture with scene-aware prompts, extended to include UAV-Human and AG-ReID.v2 datasets for comprehensive universal evaluation. Both baseline and

[1] This fallback removes online 3D extraction at inference, but not all overhead: concatenation still doubles the descriptor dimensionality from D to $2D$, and geometry extraction remains an offline preprocessing step.

Table 1: Comparison with other methods on standard holistic ReID benchmarks (Market-1501, MSMT17, and CUHK03). Each table uses a unified method column, and "-" denotes results not reported in the original papers. "*" denotes using an overlapping patch embedding layer.

Training Set	Method	Market-1501		MSMT17		CUHK03	
		R1	mAP	R1	mAP	R1	mAP
Single Scene	PCB [69]	93.8	81.6	68.2	40.4	75.3	-
	ViT-B [14]	94.0	87.6	82.8	63.6	-	-
	TransReid [21]	95.0	88.8	84.6	66.6	-	-
	MGN [71]	95.1	87.5	85.1	63.7	-	-
	ProFD [11]	95.1	90.0	-	-	-	-
	TransReid* [21]	95.2	89.5	86.2	69.4	-	-
	AAformer [104]	95.4	87.7	83.6	63.2	-	-
	AGW [87]	95.5	89.5	81.2	59.7	87.3	88.4
	FlipReID [57]	95.5	89.6	85.6	68.0	-	-
	PFD* [73]	95.5	89.7	83.8	64.4	-	-
	PartFormer [70]	95.6	89.7	85.2	68.3	-	-
	LDS [90]	95.8	90.4	86.5	67.2	-	-
	SCFP [95]	95.8	90.0	-	-	84.3	82.7
	CPHMNet [82]	95.9	89.9	82.5	61.6	-	-
	MPN [12]	96.4	90.1	83.5	62.7	-	-
	CAD-Net [40]	-	-	-	-	82.1	-
	OSNet [102]	-	-	-	-	83.8	85.4
	ABD-Net [5]	-	-	-	-	84.3	85.9
	INTACT [8]	-	-	-	-	86.4	-
	JBIM [99]	-	-	-	-	88.7	90.3
	PS-HRNet [91]	-	-	-	-	92.6	-
Multiple Scenes	AIM(R50) [86]	91.9	81.6	72.6	46.4	89.3	89.9
	ETNDNET(ViT-B) [13]	92.7	81.6	63.8	41.5	90.8	89.7
	PCB [69]	93.3	83.1	69.9	48.5	87.3	90.4
	ETNDNET(R50) [13]	93.7	84.3	75.7	50.5	90.2	89.8
	PFD [73]	94.3	86.8	82.0	63.3	94.1	95.5
	AIM(ViT-B) [86]	94.4	88.5	83.5	65.3	92.5	91.9
	TransReid [21]	95.0	89.7	85.8	69.4	92.9	93.6
	ReIDCaps [26]	95.6	90.6	86.6	69.6	93.3	94.6
Universal	Baseline (Pure 2D)	96.5	92.7	**87.5**	71.3	**96.8**	**95.9**
	Ours	**96.6**	**92.9**	87.4	**71.6**	96.6	95.6

our full model are trained under identical protocols (180 epochs, CUDA 12 environment) to ensure fair comparison. This controlled setup isolates the contribution of our reliability-gated 3D injection mechanism. Unless otherwise stated, we report raw results on nine representative benchmarks and transfer protocols in Tables 1–3. The 12-protocol aggregate in Tab. 6 further includes all four AG-ReID.v2 transfer settings together with UAV-Human.

4.2 Comparison with State-of-the-Art Methods

We benchmark our dynamically gated model against a controlled pure-2D baseline with a shared backbone and training protocol. Results in Tables 1, 2 and 4 to 6 show clearer gains in structurally challenging scenarios while maintaining competitive performance on clean benchmarks. This comparison also clarifies, in this universal setting and practice, when geometry contributes complementary evidence rather than redundant cues.

Table 2: Comparison across four ReID scenes. (a) Clothing-change (PRCC, Celeb-ReID); (b) Occlusion (Occluded-Duke) and Cross-modality (SYSU-MM01). Each subtable uses a method column, and "-" denotes results not reported in the papers.

(a) PRCC and Celeb-ReID

Method	PRCC		Celeb	
	R1	mAP	R1	mAP
Single Scene				
RCSANet [25]	31.6	31.5	55.6	11.9
MGN [71]	33.8	35.9	49.0	10.8
CASE-Net‡ [41]	39.5	-	**66.4**	18.2
PCB [69]	41.8	38.7	37.1	8.2
AFD-Net [83]	42.8	-	52.1	10.6
ViT-B [14]	44.3	57.5	59.7	13.4
TransReid [21]	45.0	57.9	58.9	14.6
RCSANet‡ [25]	50.2	48.6	65.3	17.5
LightMBN [22]	50.5	51.2	57.3	14.3
FSAM‡ [23]	54.5	-	-	-
IRANet‡ [65]	54.9	53.0	64.1	19.0
CAL [20]	55.2	55.8	-	-
CLIP3DReID [50]	**60.6**	59.3	63.1	**19.2**
ReIDCaps [26]	-	-	51.2	9.8
Multiple Scenes				
PCB [69]	44.8	57.5	10.1	1.7
ETNDNET(R50) [13]	45.5	57.2	55.8	12.7
ETNDNET(ViT-B) [13]	47.6	59.2	49.8	8.5
AIM(ViT-B) [86]	48.2	60.7	51.3	10.5
TransReid [21]	48.5	62.2	57.4	14.8
ReIDCaps [26]	50.3	64.4	59.9	17.7
AIM(R50) [86]	55.4	66.0	54.5	12.4
PFD [73]	55.5	65.8	54.0	11.0
Universal				
Baseline (Pure 2D)	56.0	68.0	60.0	16.5
Ours	59.0	**70.5**	60.4	16.8

(b) Occluded-Duke and SYSU-MM01

Method	Occ.-Duke		SYSU	
	R1	mAP	R1	mAP
Single Scene				
TransReid* [21]	66.4	59.2	-	-
FED [77]	68.1	56.4	-	-
CPHMNet [82]	68.7	61.0	-	-
PFD* [73]	69.5	61.8	-	-
ProFD [11]	70.8	62.8	-	-
SCFP [95]	70.9	63.0	-	-
SCING [81]	71.1	63.4	-	-
PAB-ReID [7]	72.6	63.5	-	-
MA-MSA [28]	73.3	62.9	-	-
HAT [88]	-	-	55.3	53.9
HC [106]	-	-	57.0	55.0
VT-ReID [51]	-	-	61.7	57.5
CM-NAS [16]	-	-	62.0	60.0
Multiple Scenes				
ETNDNET(ViT-B) [13]	44.8	38.4	51.4	43.4
PCB [69]	46.7	39.0	52.7	48.3
AIM(R50) [86]	52.3	42.9	58.2	56.4
ETNDNET(R50) [13]	58.2	48.7	55.1	50.0
PFD [73]	59.5	52.0	62.7	62.0
AIM(ViT-B) [86]	62.9	54.2	60.6	60.2
TransReid [21]	64.2	56.6	58.8	59.5
ReIDCaps [26]	66.7	58.8	63.2	59.9
UFFM+AMC [2]	68.9	61.9	-	-
Universal				
Baseline (Pure 2D)	73.9	65.5	63.1	64.3
Ours	**74.8**	**65.8**	**64.3**	**65.4**

Note: "‡" denotes contour-based methods; "*" means overlapping patch embedding.

Table 3: Performance comparison on UAV-Human and AG-ReID.v2. A→A denotes the aerial-to-aerial protocol on UAV-Human. A→C, A→W, C→A, and W→A are AG-ReID.v2 protocols, where A, C, and W denote aerial, CCTV, and wearable views. "-" denotes results not reported in the papers.

Method	A→A		A→C		A→W		C→A		W→A	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
AG-ReIDv2 [56]	**72.75**	71.47	88.77	80.72	93.62	84.85	87.86	78.51	88.61	80.11
SeCap [72]	-	-	88.12	80.84	91.44	84.01	88.24	79.99	87.56	80.15
GSAlign [36]	-	-	87.86	81.38	90.63	83.98	88.02	81.05	87.31	80.90
Baseline (Pure 2D)	71.4	**73.2**	92.3	88.1	**94.0**	**90.2**	**89.5**	**83.9**	**90.7**	**85.7**
Ours	71.4	73.1	**92.6**	**88.8**	93.7	90.1	88.9	**83.9**	90.5	85.6

Prevention of Negative Transfer on Clean Domains. As shown in Table 1, on clean domains where texture cues are reliable, our method remains close to the baseline: Market-1501 (96.6% vs. 96.5%), MSMT17 (87.4% vs. 87.5%

Table 4: Ablation Study 1: Multi-Modal Fusion Strategy Analysis on core benchmarks. Rigid 3D fusion actively restricts performance, while Dynamic Gating successfully bridges the modality gap.

Method	Market		MSMT		CUHK		PRCC		Celeb		Occ.-Duke		SYSU	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Baseline (Pure 2D)	96.5	92.7	**87.5**	71.3	**96.8**	**95.9**	56.0	68.0	60.0	16.5	73.9	65.5	63.1	64.3
+ Naive 3D	**96.6**	**93.0**	87.1	71.5	96.2	95.6	55.0	67.1	59.9	16.3	73.7	65.4	**64.6**	**65.4**
+ RG-3D (Ours)	**96.6**	92.9	87.4	**71.6**	96.6	95.6	**59.0**	**70.5**	**60.4**	**16.8**	**74.8**	**65.8**	64.3	**65.4**

R1, 71.6% vs. 71.3% mAP), and CUHK03 (96.6% vs. 96.8%). This matches our conditional-use hypothesis: in holistic settings, many queries are already resolved by appearance cues, leaving limited room for additional geometry. MSMT17 still gains +0.3% mAP, while the slight CUHK03 drop shows that geometry is not an always-on benefit. This restrained behavior is consistent with the low gate activation observed in standard holistic scenarios, where the model largely falls back to visual evidence. Instead, the gate mainly helps when appearance becomes ambiguous while avoiding severe negative transfer on clean domains, which is important for universal deployment in practice.

Comparison with State-of-the-Art Methods. Our method stays competitive across scenarios. On clothing-change (PRCC: 59.0% R1, +3.0%; Celeb-ReID: 60.4% R1, +0.4%), occlusion (Occluded-Duke: 74.8% R1, +0.9%), and cross-modality (SYSU-MM01: 64.3% R1, +1.2%), it improves the settings where appearance cues are less reliable. In UAV and wild scenarios, it matches the baseline on UAV-Human Rank-1 (71.4%), improves AG-ReID.v2 A→C by +0.3% Rank-1, and remains competitive on the other transfer protocols.

4.3 Extensive Ablation Studies

To isolate more precisely where the gains come from, we ablate both fusion strategy and structural input composition.

Analysis of the Gating Mechanism. Our ablation in Table 4 verifies the necessity of the Reliability-Aware Gate. Unfiltered structural fusion (Naive 3D Concat) fails to provide stable gains, slightly improving some clean-domain mAP values while regressing on several structurally challenging settings. Relative to the Pure 2D baseline, our dynamically learned reliability gate improves PRCC by +3.0% Rank-1, Occluded-Duke by +0.9%, and SYSU-MM01 by +1.2%, while remaining comparable on standard benchmarks. Its value is especially evident in stability: across the seven reported benchmarks, our gated fusion never underperforms the baseline by more than 0.3% on either metric, whereas Naive Concat shows larger variance (e.g., PRCC Rank-1 drops 1.0%). This contrast shows that the key benefit is not simply adding structural capacity, but controlling when geometric evidence participates in the final representation under heterogeneous deployment conditions, thereby avoiding catastrophic negative transfer. Detailed gate activation analysis is provided in Sec. 4.4.

Table 5: Ablation Study 2: Decoupled SMPL parametric modeling.

Method	Market		MSMT		CUHK		PRCC		Celeb		Occ.-Duke		SYSU	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Black-box SMPL†	95.7	92.0	85.4	69.2	96.2	95.0	51.3	64.5	59.5	16.7	66.7	59.1	63.2	**65.8**
Baseline (Pure 2D)	96.5	92.7	87.5	71.3	**96.8**	**95.9**	56.0	68.0	60.0	16.5	73.9	65.5	63.1	64.3
Global Params	96.6	92.9	**87.6**	71.5	96.7	95.7	57.2	69.0	60.2	**16.9**	73.6	65.5	64.8	65.3
Full Topology*	96.6	92.9	87.4	**71.6**	96.6	95.6	**59.0**	**70.5**	**60.4**	16.8	**74.8**	**65.8**	64.3	65.4
Local Joints	**96.8**	**93.0**	87.5	**71.6**	96.6	95.7	56.9	68.8	60.1	16.7	73.4	65.2	**65.0**	65.4

Notation: Black-box=$\mathbf{s}$, Global=$\mathbf{s}_{global}$, Local=$\mathbf{s}_{joint}$, Full=$\mathbf{s}_{global} + \mathbf{s}_{joint}$.
† epoch 120; others epoch 180. * joint-token pooling only.

Table 6: Fusion Design Evolution: Comparison against stronger 3D fusion alternatives. Averages are computed over 12 protocols, namely the seven benchmark subset shown in Tables 1–5, plus UAV-Human (A→A) and the four AG-ReID.v2 transfer protocols. The scalar reliability gate is the only 3D-augmented variant that surpasses the pure-2D baseline, achieving Pareto-optimal performance-complexity trade-off.

Method	Avg mAP	Avg Rank-1	Extra Params	Seq. Cost
HyperNet Prompt†	72.1	78.5	∼0.8M	$+K$ tokens
3D-SIE (Token Injection)	72.7	76.2	∼1.2M	+24 tokens
MoE-Expert (4-Expert Routing)	73.2	77.1	∼10M	none
Pure 2D Baseline	74.6	81.0	none	none
Reliability-Gated (Ours)	**75.0**	**81.4**	∼0.3M	none

† HyperNet Prompt evaluated at epoch 60 due to training instability in subsequent epochs.

Necessity of Kinematic-Aware Decoupling. Our comparison in Table 5 isolates the impact of different SMPL modeling strategies. Directly processing the un-decoupled 82-D SMPL vector (Black-box SMPL) induces severe instability characterized by gradient explosion and loss divergence, requiring early termination at epoch 120. This validates our hypothesis: neglecting explicit kinematic parameter semantics entangles invariant factors with volatile variables, destabilizing training. In contrast, explicitly isolating global parameters ($\mathbf{s}_{global}$) and local joint streams ($\mathbf{s}_{joint}$) preserves spatial semantics. Our full topology variant provides the best overall cross-scenario trade-off, while SYSU-MM01 remains a notable exception where local joints achieve slightly higher Rank-1. This suggests that semantically constrained modeling is essential for effective 3D injection, but cross-modality settings can favor the more sensor-stable local-joint representation. This also clarifies the role of the global token. Rather than serving as a direct identity descriptor, it acts as contextual support for joint refinement, allowing global body cues to inform local structure while reducing the risk of view- or sensor-dependent bias in the final representation.

Fusion Design Evolution. To validate the scalar gate, we compared it with three stronger 3D fusion alternatives: MoE-Expert, 3D-SIE, and HyperNet Prompt. Results in Tab. 6 show that all three underperform the pure-2D baseline on the 12-protocol average: MoE-Expert drops 1.4% mAP and 3.9% Rank-1, 3D-SIE drops 1.9% mAP and 4.8% Rank-1, and HyperNet Prompt is unstable,

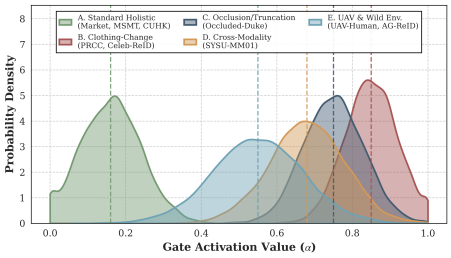

Fig. 3: Dynamic Distribution of Reliability Gate (α) across different representative ReID scenario groups.

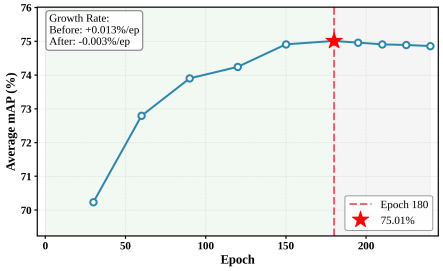

Fig. 4: Training Convergence Analysis showing average mAP trends over epochs during universal joint training.

reaching only 72.1% avg mAP at epoch 60. In contrast, Reliability-Gated is the only 3D-augmented variant that improves over the baseline (+0.4% avg mAP, +0.4% avg Rank-1) with only ∼0.3M extra parameters. These stronger alternatives inject geometry too early or too aggressively, forcing noisy structural cues to interact with visual tokens before reliability is assessed. In universal training, where monocular 3D quality varies sharply across datasets and samples, our design instead delays geometric intervention to the final representation stage and modulates structural contribution through cross-modal consistency. The 12-protocol average covers CUHK03, Market-1501, MSMT17, Celeb-ReID, PRCC, Occluded-Duke, SYSU-MM01, UAV-Human (A→A), and all four AG-ReID.v2 transfer protocols used here for evaluation.

4.4 Analysis and Visualizations

Quantitative Validation of Gate Activation (α). To better understand gate behavior, we analyze the activation distribution of the scalar α over all query samples at test time. Activation patterns in Fig. 3 vary clearly across scenario groups. For standard holistic scenarios (Group A: Market-1501, MSMT17, CUHK03), the mean α is approximately 0.15 ± 0.05, down-weighting 3D cues and relying on 2D textures. Conversely, for clothing-change scenarios (Group B: PRCC and Celeb-ReID), the mean α increases to 0.82 ± 0.10, indicating stronger reliance on geometric posture when appearance cues are unreliable. Occlusion scenarios (Group C: Occluded-Duke, $\mu = 0.75$) and cross-modality scenarios (Group D: SYSU-MM01, $\mu = 0.68$) also show elevated 3D injection, while UAV and wild scenarios (Group E: UAV-Human and AG-ReID.v2, $\mu = 0.55$) remain intermediate, suggesting a more balanced fusion regime under large viewpoint variation. Overall, the learned gate adaptively weights geometric priors by visual-feature reliability across challenging conditions. It acquires this behavior through end-to-end supervision alone, and future work can strengthen the mechanism further by adding explicit geometric quality proxies (*e.g.* 2D keypoint confidence or SMPL fitting scores) as auxiliary reliability signals.

Training Convergence and Stability. The training curve in Figure 4 shows rapid ascent followed by steady convergence, supporting the stability of the proposed reliability-gated optimization. The smooth plateau suggests that scalar reliability modulation avoids the volatility often introduced by more entangled multimodal coupling. This behavior matches our late residual design, which limits cross-stream interference and aligns with the ablations where stronger early-fusion variants underperform or are harder to optimize. Notably, no late-stage oscillation appears after the main performance rise, indicating that the gate settles into a stable cross-modal weighting regime rather than repeatedly over-correcting noisy geometry. This trend is consistent with the clean-domain fallback behavior observed elsewhere, where visual features remain the dominant anchor unless structural cues provide reliably complementary evidence.

5 Conclusion

In this paper, we integrated monocular 3D geometric priors into Universal person re-identification through a Reliability-Aware Geometric Injection framework. By decoupling geometric extraction from dynamic utilization, our method explicitly models kinematic joint topology and structure. Coupled with a consistency-aware gating mechanism, the proposed dual-stream residual fusion adaptively leverages geometric cues while maintaining competitive performance on standard holistic benchmarks across diverse deployment scenarios.

Extensive experiments validate this conditional-use perspective, with clear gains on PRCC (+3.0% Rank-1), Occluded-Duke (+0.9%), and SYSU-MM01 (+1.2%), while preserving strong results on standard holistic settings. These results suggest that geometric priors are most valuable when appearance evidence is degraded, and that reliability-aware fallback is essential for avoiding negative transfer in universal deployment. More broadly, our findings indicate that auxiliary geometry should be treated as conditional structural evidence rather than mandatory parallel input, especially in heterogeneous real-world environments where 3D quality can vary substantially across domains and samples.

This interpretation helps reconcile why geometry is highly effective for occlusion, clothing-change, and cross-modality settings, yet should remain selectively suppressed in already well-resolved holistic cases. It also reinforces a broader practical lesson for universal deployment: auxiliary modalities should contribute conditionally, in proportion to their reliability, rather than being fused as uniformly trustworthy signals in practice. Future work will correlate the learned gate with explicit geometric quality metrics and explore temporal geometric aggregation for video-based ReID.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. T2541022 and the China Postdoctoral Science Foundation - Hubei Joint Support Program under Grant No. 2025T053HB.

References

1. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European conference on computer vision. pp. 561–578. Springer (2016)
2. Che, Q.H., Nguyen, L.C., Luu, D.T., Nguyen, V.T.: Enhancing person re-identification via uncertainty feature fusion method and auto-weighted measure combination. Knowledge-Based Systems **307**, 112737 (2025)
3. Chen, C., Ye, M., Jiang, D.: Towards modality-agnostic person re-identification with descriptive query. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15128–15137 (2023)
4. Chen, J., Jiang, X., Wang, F., Zhang, J., Zheng, F., Sun, X., Zheng, W.S.: Learning 3d shape feature for texture-insensitive person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8146–8155 (2021)
5. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8351–8361 (2019)
6. 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. pp. 15050–15061 (2023)
7. Chen, Z., Ge, Y.: Part-attention based model make occluded person re-identification stronger. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
8. Cheng, 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. pp. 2605–2615 (2020)
9. Cho, J., Youwang, K., Oh, T.H.: Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In: European Conference on Computer Vision. pp. 342–359. Springer (2022)
10. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10257–10266 (2020)
11. Cui, C., Huang, S., Song, W., Ding, P., Zhang, M., Wang, D.: Profd: Prompt-guided feature disentangling for occluded person re-identification. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1583–1592 (2024)
12. Ding, C., Wang, K., Wang, P., Tao, D.: Multi-task learning with coarse priors for robust part-aware person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3), 1474–1488 (2020)
13. Dong, N., Zhang, L., Yan, S., Tang, H., Tang, J.: Erasing, transforming, and noising defense network for occluded person re-identification. IEEE Transactions on Circuits and Systems for Video Technology **34**(6), 4458–4472 (2023)
14. 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. arXiv preprint arXiv:2010.11929 (2020)
15. Feng, J., Wu, A., Zheng, W.S.: Shape-erased feature learning for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22752–22761 (2023)

16. Fu, C., Hu, Y., Wu, X., Shi, H., Mei, T., He, R.: Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11823–11832 (2021)
17. Gao, S., Wang, J., Lu, H., Liu, Z.: 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 (2020)
18. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: International conference on learning representations (2018)
19. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14783–14794 (2023)
20. Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., Chen, X.: Clothes-changing person re-identification with rgb modality only. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1060–1069 (2022)
21. 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. pp. 15013–15022 (2021)
22. Herzog, F., Ji, X., Teepe, T., Hörmann, S., Gilg, J., Rigoll, G.: Lightweight multi-branch network for person re-identification. In: 2021 IEEE International conference on image processing (ICIP). pp. 1129–1133. IEEE (2021)
23. Hong, P., Wu, T., Wu, A., Han, X., Zheng, W.S.: Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10513–10522 (2021)
24. Huang, Y., Wu, Q., Xu, J., Zhong, Y.: Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2019)
25. Huang, Y., Wu, Q., Xu, J., Zhong, Y., Zhang, Z.: Clothing status awareness for long-term person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11895–11904 (2021)
26. 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 Transactions on Circuits and Systems for Video Technology **30**(10), 3459–3471 (2019)
27. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5579–5588 (2020)
28. Jiang, X., Yuan, X., Yang, X.: Multi-aligned and multi-scale augmentation for occluded person re-identification. Sensors **25**(19), 6210 (2025)
29. Jin, X., He, T., Zheng, K., Yin, Z., Shen, X., Huang, Z., Feng, R., Huang, J., Chen, Z., Hua, X.S.: Cloth-changing person re-identification from a single image with gait prediction and regularization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14278–14287 (2022)
30. Joseph, A., Peleg, S.: Clothes-changing person re-identification based on skeleton dynamics. arXiv preprint arXiv:2503.10759 (2025)

31. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018)
32. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
33. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11127–11137 (2021)
34. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2252–2261 (2019)
35. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11605–11614 (2021)
36. Li, Q., Li, J., Zhang, Y., Tan, L., Chen, J., Ji, J.: Gsalign: Geometric and semantic alignment network for aerial-ground person re-identification. arXiv preprint arXiv:2510.22268 (2025)
37. 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. vol. 37, pp. 1405–1413 (2023)
38. Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., Li, Z.: Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16266–16275 (2021)
39. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 152–159 (2014)
40. Li, Y.J., Chen, Y.C., Lin, Y.Y., Du, X., Wang, Y.C.F.: Recover and identify: A generative dual model for cross-resolution person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8090–8099 (2019)
41. Li, Y.J., Weng, X., Kitani, K.M.: Learning shape representations for person re-identification under clothing change. In: Proceedings of the IEEE/CVF Winter conference on applications of computer vision. pp. 2432–2441 (2021)
42. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: Occluded person re-identification with part-aware transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2898–2907 (2021)
43. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: European Conference on Computer Vision. pp. 590–606. Springer (2022)
44. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21159–21168 (2023)
45. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1954–1963 (2021)
46. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12939–12948 (2021)
47. Liu, F., Ye, M., Du, B.: Dual level adaptive weighting for cloth-changing person re-identification. IEEE Transactions on Image Processing **32**, 5075–5086 (2023)

48. Liu, F., Ye, M., Du, B.: Cloth-aware augmentation for cloth-generalized person re-identification. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 4053–4062 (2024)
49. Liu, F., Kim, M., Gu, Z., Jain, A., Liu, X.: Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19617–19626 (2023)
50. Liu, F., Kim, M., Ren, Z., Liu, X.: Distilling clip with dual guidance for learning discriminative human body shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 256–266 (2024)
51. Liu, H., Tan, X., Zhou, X.: Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. IEEE Transactions on Multimedia **23**, 4414–4425 (2020)
52. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248:1–248:16 (2015). https://doi.org/10.1145/2816795.2818013
53. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
54. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 542–551 (2019)
55. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: European Conference on Computer Vision. pp. 752–768. Springer (2020)
56. Nguyen, H., Nguyen, K., Sridharan, S., Fookes, C.: Ag-reid. v2: Bridging aerial and ground views for person re-identification. IEEE Transactions on Information Forensics and Security **19**, 2896–2908 (2024)
57. Ni, X., Rahtu, E.: Flipreid: closing the gap between training and inference in person re-identification. In: 2021 9th European Workshop on Visual Information Processing (EUVIP). pp. 1–6. IEEE (2021)
58. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV). pp. 484–494. IEEE (2018)
59. Pang, Z., Wang, J., Zhao, L., Wang, C.: Identity-clothing similarity modeling for unsupervised clothing change person re-identification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19251–19260 (2025)
60. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE signal processing magazine **34**(6), 96–108 (2017)
61. Rao, H., Hu, X., Cheng, J., Hu, B.: Sm-sge: A self-supervised multi-scale skeleton graph encoding framework for person re-identification. In: Proceedings of the 29th ACM international conference on Multimedia. pp. 1812–1820 (2021)
62. Rao, H., Miao, C.: Transg: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22118–22128 (2023)
63. Ren, K., Zhang, L.: Implicit discriminative knowledge learning for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 393–402 (2024)

64. Ren, Y., Wang, Y., Tan, S., Chen, Y., Yang, J.: Person re-identification in 3d space: A {WiFi} vision-based approach. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 5217–5234 (2023)
65. Shi, W., Liu, H., Liu, M.: Iranet: Identity-relevance aware representation for cloth-changing person re-identification. Image and vision computing **117**, 104335 (2022)
66. Somers, V., Alahi, A., Vleeschouwer, C.D.: Keypoint promptable re-identification. In: European Conference on Computer Vision. pp. 216–233. Springer (2024)
67. Somers, V., De Vleeschouwer, C., Alahi, A.: Body part-based representation learning for occluded person re-identification. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1613–1623 (2023)
68. Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 608–617 (2019)
69. 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). pp. 480–496 (2018)
70. Tan, L., Dai, P., Chen, J., Cao, L., Wu, Y., Ji, R.: Partformer: Awakening latent diverse representation from vision transformer for object re-identification. arXiv preprint arXiv:2408.16684 (2024)
71. 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. pp. 274–282 (2018)
72. Wang, S., Wang, Y., Wu, R., Jiao, B., Wang, W., Wang, P.: Secap: self-calibrating and adaptive prompts for cross-view person re-identification in aerial-ground networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22119–22128 (2025)
73. Wang, T., Liu, H., Song, P., Guo, T., Shi, W.: Pose-guided feature disentangling for occluded person re-identification based on transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2540–2549 (2022)
74. Wang, Y., Liao, S., Shao, L.: Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In: Proceedings of the 28th ACM international conference on multimedia. pp. 3422–3430 (2020)
75. Wang, Y., Chen, Z., Wu, F., Wang, G.: Person re-identification with cascaded pairwise convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1470–1478 (2018)
76. Wang, Y., Yu, H., Yan, Y., Song, S., Liu, B., Lu, Y.: Exploring shape embedding for cloth-changing person re-identification via 2d-3d correspondences. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7121–7130 (2023)
77. 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. pp. 4754–4763 (2022)
78. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 79–88 (2018)
79. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE international conference on computer vision. pp. 5380–5389 (2017)

80. Wu, L., Wang, Y., Shao, L., Wang, M.: 3-d personvlad: Learning deep global representations for video-based person reidentification. IEEE transactions on neural networks and learning systems **30**(11), 3347–3359 (2019)
81. Xie, Y., Cheng, Y., Wu, J., Zhang, H., Zhou, Y., Han, S.: Scing: Towards more efficient and robust person re-identification through selective cross-modal prompt tuning. arXiv preprint arXiv:2507.00506 (2025)
82. Xu, H., Gao, R.: A dual-branch pedestrian re-identification method cphmnet based on multi-dimensional feature fusion and integrated pose estimation. Scientific Reports **15**(1), 34912 (2025)
83. Xu, W., Liu, H., Shi, W., Miao, Z., Lu, Z., Chen, F.: Adversarial feature disentanglement for long-term person re-identification. In: IJCAI. pp. 1201–1207 (2021)
84. Yan, C., Pang, G., Jiao, J., Bai, X., Feng, X., Shen, C.: Occluded person re-identification with single-scale global representations. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11875–11884 (2021)
85. Yang, Q., Wu, A., Zheng, W.S.: Person re-identification by contour sketch under moderate clothing change. IEEE transactions on pattern analysis and machine intelligence **43**(6), 2029–2046 (2019)
86. Yang, Z., Lin, M., Zhong, X., Wu, Y., Wang, Z.: Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1472–1481 (2023)
87. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence **44**(6), 2872–2893 (2021)
88. Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Transactions on Information Forensics and Security **16**, 728–739 (2020)
89. Yu, Z., Tiwari, P., Hou, L., Li, L., Li, W., Jiang, L., Ning, X.: Mv-reid: 3d multi-view transformation network for occluded person re-identification. Knowledge-Based Systems **283**, 111200 (2024)
90. Zang, X., Li, G., Gao, W., Shu, X.: Learning to disentangle scenes for person re-identification. Image and Vision Computing **116**, 104330 (2021)
91. Zhang, G., Ge, Y., Dong, Z., Wang, H., Zheng, Y., Chen, S.: Deep high-resolution representation learning for cross-resolution person re-identification. IEEE Transactions on Image processing **30**, 8913–8925 (2021)
92. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11446–11456 (2021)
93. Zhang, J., Wu, Y., Jiang, H.: Survey on monocular metric depth estimation. Computers **14**(11), 502 (2025)
94. Zhang, T., Xie, L., Wei, L., Zhuang, Z., Zhang, Y., Li, B., Tian, Q.: Unrealperson: An adaptive pipeline towards costless person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11506–11515 (2021)
95. Zhao, C., Qin, Y., Zhang, B., Zhao, Y., Wu, B.: An end-to-end occluded person re-identification network with smoothing corrupted feature prediction. Artificial Intelligence Review **58**(2), 53 (2024)
96. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose-invariant embedding for deep person re-identification. IEEE transactions on image processing **28**(9), 4500–4509 (2019)

97. 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. pp. 1116–1124 (2015)
98. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
99. Zheng, W.S., Hong, J., Jiao, J., Wu, A., Zhu, X., Gong, S., Qin, J., Lai, J.: Joint bilateral-resolution identity modeling for cross-resolution person re-identification. International Journal of Computer Vision **130**(1), 136–156 (2022)
100. Zheng, W.S., Yan, J., Peng, Y.X.: A versatile framework for multi-scene person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 1362–1380 (2024)
101. Zheng, Z., Wang, X., Zheng, N., Yang, Y.: Parameter-efficient person re-identification in the 3d space. IEEE Transactions on Neural Networks and Learning Systems **35**(6), 7534–7547 (2022)
102. 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. pp. 3702–3712 (2019)
103. Zhu, K., Guo, H., Liu, Z., Tang, M., Wang, J.: Identity-guided human semantic parsing for person re-identification. In: European conference on computer vision. pp. 346–363. Springer (2020)
104. Zhu, K., Guo, H., Zhang, S., Wang, Y., Liu, J., Wang, J., Tang, M.: Aaformer: Auto-aligned transformer for person re-identification. IEEE transactions on neural networks and learning systems **35**(12), 17307–17317 (2023)
105. Zhu, Y., Li, A., Tang, Y., Zhao, W., Zhou, J., Lu, J.: Dpmesh: Exploiting diffusion prior for occluded human mesh recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1101–1110 (2024)
106. Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X., Tao, D.: Hetero-center loss for cross-modality person re-identification. Neurocomputing **386**, 97–109 (2020)