electronics

MDPI

Review

Causality and "In-the-Wild" Video-Based Person Re-Identification: A Survey

Md Rashidunnabi [1,2,*], Kailash Hambarde [1] and Hugo Proença [1,2]

[1] Department of Computer Science, University of Beira Interior, 6201-001 Covilhã, Portugal; N1947@ubi.pt (K.H.); hugomcp@ubi.pt (H.P.)
[2] IT: Instituto de Telecomunicações, University of Beira Interior, 6201-001 Covilhã, Portugal
* Correspondence: md.rashidunnabi@ubi.pt

Abstract

Video-based person re-identification (re-identification) remains underused in real-world deployments, despite impressive benchmark performance. Most existing models rely on superficial correlations—such as clothing, background, or lighting—that fail to generalize across domains, viewpoints, and temporal variations. This study examines the emerging role of causal reasoning as a principled alternative to traditional correlation-based approaches in video-based re-identification. We provide a structured and critical analysis of methods that leverage structural causal models (SCMs), interventions, and counterfactual reasoning to isolate identity-specific features from confounding factors. This study is organized around a novel taxonomy of causal re-identification methods, spanning generative disentanglement, domain-invariant modeling, and causal transformers. We review current evaluation metrics and introduce causal-specific robustness measures. In addition, we assess the practical challenges—scalability, fairness, interpretability, and privacy—that must be addressed for real-world adoption. Finally, we identify open problems and outline future research directions that integrate causal modeling with efficient architectures and self-supervised learning. This study aims to establish a coherent foundation for causal video-based person re-identification and catalyze the next phase of research in this rapidly evolving domain.

Keywords: video-based person re-identification; causal inference; structural causal models; counterfactual reasoning; transformer architectures

check for updates

Academic Editors: Reza K. Amineh and Xin Geng

Received: 26 May 2025
Revised: 22 June 2025
Accepted: 28 June 2025
Published: 1 July 2025

Citation: Rashidunnabi, M.; Hambarde, K.; Proença, H. Causality and "In-the-Wild" Video-Based Person Re-Identification: A Survey. *Electronics* **2025**, *14*, 2669. https://doi.org/10.3390/electronics14132669

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Video-based person re-identification (re-identification) is a critical task in computer vision, with applications in surveillance, smart cities, and forensics [1]. Unlike image-based re-identification, which relies on static appearance cues, video-based methods leverage temporal sequences—capturing motion, gait, and behavioral dynamics—to match individuals across non-overlapping camera views [2,3]. This added temporal dimension provides richer identity signals, particularly in unconstrained environments where single-frame models often fail. Here, 'in-the-wild' refers to unconstrained, real-world surveillance scenarios that exhibit large variations in illumination, viewpoint, occlusion, weather, and attire.

Despite substantial progress, most video-based re-identification systems remain underused in real-world conditions. Benchmark-leading models degrade sharply when exposed to domain shifts, occlusions, lighting changes, or clothing variations [4,5]. The root cause is methodological: these models are correlation-driven, trained to optimize performance

on tightly curated datasets by exploiting superficial cues—such as clothing color or background texture—that do not generalize to real deployments [6–8]. This leads to fragile identity representations that collapse under distribution shift [9].

To overcome these limitations, causal inference offers a fundamentally different paradigm. Rather than modeling statistical associations between visual input and identity, causal methods aim to isolate the true generative factors of identity—such as body shape, gait, and motion patterns—while explicitly controlling for confounding variables like clothing, background, and viewpoint [10–12]. Structural Causal Models (SCMs), counterfactual reasoning, and interventional training frameworks provide the tools to enforce this separation, enabling models that are more robust, generalizable, and interpretable [5,13–15].

As illustrated in Figure 1, a robust re-identification system must ignore nuisance variation and preserve consistent identity representations across dramatic appearance shifts. Figure 2 highlights the practical benefits of causal disentanglement for deployment robustness. Causal methods explicitly model this requirement by intervening in non-identity attributes and learning representations invariant to them. This shift enables models to resist shortcut learning and focus on stable identity features that remain consistent across environments.



Figure 1. Why video-based person re-identification is hard. The same individual appears under six nuisance factors—viewpoint, lighting, rain blur, pose, clothing change, and accessory occlusion—illustrating the need for causal disentanglement rather than correlation-driven learning.

This study provides a structured and critical overview of causal approaches in video-based person re-identification. Our contributions are as follows:

- We provide a comprehensive taxonomy of causal methods in re-identification, covering structural modeling, interventional training, adversarial disentanglement, and counterfactual evaluation;
- We review state-of-the-art causal re-identification models (e.g., DIR-ReID, IS-GAN, UCT) and analyze their performance across real-world challenges such as clothing change, domain shift, and multi-modality;

- We propose a unified causal framework for reasoning about identity, confounders, and interventions in re-identification pipelines;
- We discuss emerging causal evaluation metrics, interpretability tools, and benchmark gaps that must be addressed for widespread adoption;
- We identify open problems and outline future research directions at the intersection of causality, efficiency, privacy, and fairness in real-world re-identification systems.

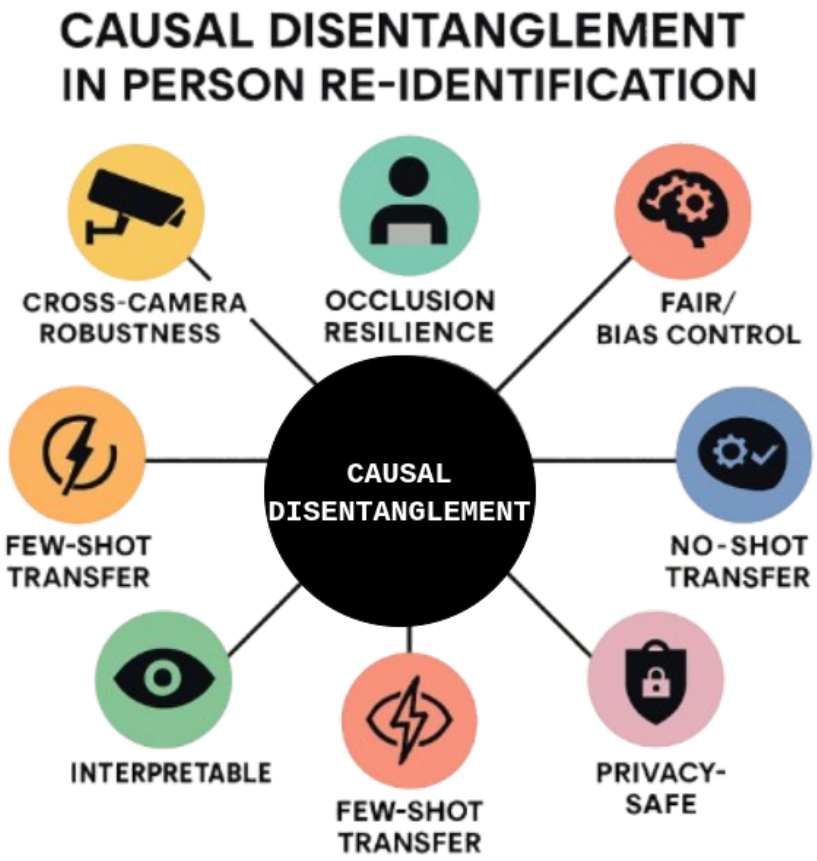


Figure 2. Benefits of Causal Disentanglement in Video-Based Person re-identification. Causal reasoning improves cross-domain robustness, occlusion resilience, fairness, privacy, and interpretability—key for real-world re-identification systems.

The remainder of this paper is structured as follows. Section 2 reviews the foundations and limitations of conventional re-identification approaches. Section 3 introduces causal inference and formalizes its role in disentangling identity from confounders. Section 4 presents a comprehensive taxonomy of causal video-based person re-identification methods. Section 5 surveys state-of-the-art causal re-identification models. Section 6 details causal disentanglement strategies in practice. Section 7 discusses the current state and practical considerations of causal video-based person re-identification and open challenges. Section 8 presents future directions. Finally, Section 9 summarizes key insights and calls for a paradigm shift from correlation to causation in video-based person re-identification.

2. Fundamentals of Person Re-Identification

2.1. Overview of Video-Based Person Re-Identification

Video-based person re-identification (re-identification) focuses on matching individuals across different camera views using sequences of video frames, called tracklets. Unlike single-image re-identification, which relies on appearance cues, video-based methods leverage both spatial (appearance) and temporal (motion) information. This combination is

crucial for distinguishing individuals, especially when appearance alone is unreliable due to variations in viewpoint, illumination, or clothing [2,4]. Motion information, such as gait and temporal dynamics, plays a significant role in video-based re-identification. While appearance-based features like clothing color or body shape are useful, they can change due to factors like lighting, posture, or occlusion. In contrast, motion patterns remain relatively stable and can help maintain identity consistency across camera views.

The typical video-based person re-identification pipeline, as shown in Figure 3, involves frame-level feature extraction, temporal modeling using RNNs or 3D CNNs, and sequence aggregation to generate a fixed-length identity representation. These methods allow for the capture of both appearance and motion features, which is essential for matching tracklets across non-overlapping camera views [3,16,17].

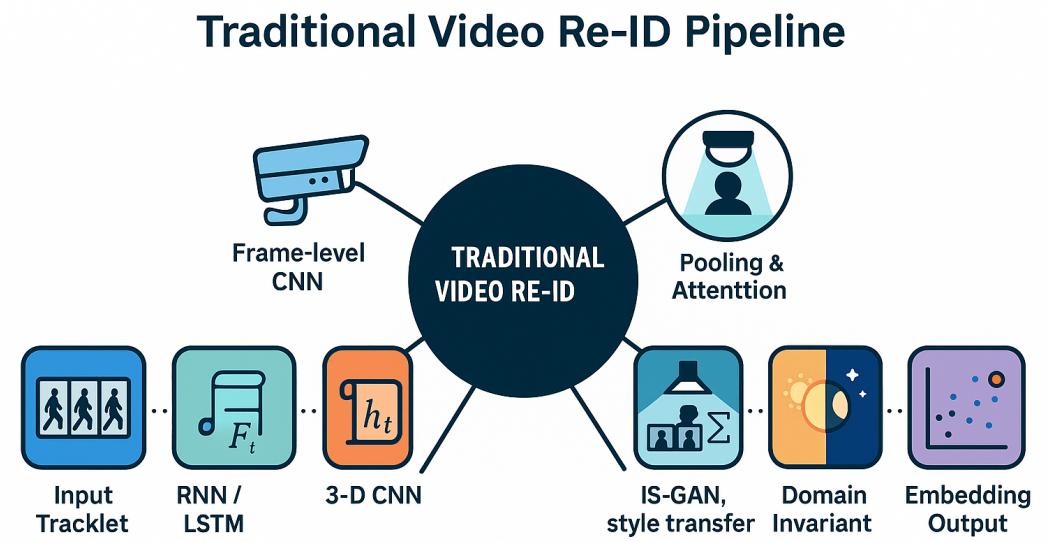


Figure 3. Traditional video-based person re-identification pipeline. The diagram summarizes classical modules—frame-level convolutional neural networks (CNNs), temporal modeling (recurrent neural networks (RNNs)/long short-term memory (LSTM)/3-D CNN), pooling–attention, generative augmentation, and domain-invariant learning—that transform a tracklet into a fixed-length identity embedding.

2.2. Challenges in Video-Based Re-Identification

Video-based person re-identification (re-identification) introduces several complexities compared to single-image person re-identification due to the dynamic nature of video data. Key challenges include the following:

Occlusions. In video sequences, individuals are often partially obscured by other objects or people, causing missing identity features. These occlusions can significantly hinder the model's ability to match tracklets across non-overlapping camera views, leading to errors in identity classification.

Viewpoint Variations. Viewpoint changes occur when individuals are captured by cameras positioned at different angles. This results in variations in appearance, as features like body shape and face may look different from different viewpoints. Video-based methods need to account for these changes, typically by utilizing temporal information such as gait and motion patterns, which remain stable across camera views [4,5].

Lighting Variations. Lighting shifts, such as day-to-night or artificial lighting changes, can cause significant color and texture changes in appearance. This can distort visual features like clothing or skin tone, leading to performance degradation in traditional appearance-based methods. Temporal modeling and domain-invariant learning techniques help mitigate these lighting-induced discrepancies [5,6].

Environmental Factors. Additional environmental factors, such as weather conditions (e.g., rain or fog), background clutter, and scene distractions, can introduce noise into the feature extraction process, further complicating identity matching. Video-based person re-identification systems must be robust to these variations, isolating true identity features from contextual distractions [4,5].

These challenges—occlusions, viewpoint variations, lighting changes, and environmental factors—necessitate video-based person re-identification systems that can robustly handle dynamic conditions. Models must be designed to focus on identity-specific features and incorporate temporal information to account for these complexities.

2.3. Traditional Approaches and Their Limitations

Traditional video-based person re-identification (re-identification) methods follow a sequential pipeline: frame-level feature extraction using convolutional neural networks (CNNs), temporal modeling via recurrent neural networks (RNNs, including long short-term memory networks, LSTMs) or 3D CNNs, and sequence aggregation through pooling or attention mechanisms [2,3]. Frame features $F_t = \text{CNN}(x_t; \theta_{\text{cnn}})$ are temporally aggregated using RNN updates $h_t = g(W_h h_{t-1} + W_x F_t + b_h)$ and attention weights $\alpha_t = \frac{\exp(W_a h_t)}{\sum_{t=1}^{T} \exp(W_a h_t)}$ to produce tracklet-level identity embeddings [17]. Recent methods incorporate generative augmentation (Identity Shuffle Generative Adversarial Network, IS-GAN) and domain-invariant learning (Domain-Invariant Representation Learning for re-identification, DIR-re-identification) to improve robustness [5,16].

Critical Limitations vs. Causal Approaches: The fundamental weakness of traditional methods lies in their correlation-driven learning paradigm. These approaches optimize statistical associations between visual inputs and identity labels without distinguishing between genuine identity characteristics and spurious environmental correlations [5,10]. This leads to three key failure modes that causal methods directly address:

(1) **Spurious Correlation Dependence:** Traditional models conflate identity-specific features (gait, body structure) with confounding factors (clothing, background, lighting), causing performance degradation under domain shifts [4,6]. Causal alternative: Structural causal models (SCMs) explicitly separate identity factors from confounders through interventional training, ensuring robust identity representations [5,12].

(2) **Lack of Invariance Guarantees:** RNN-based temporal modeling and attention mechanisms fail to provide theoretical guarantees about feature invariance across environmental changes [18,19]. Causal alternative: Counterfactual reasoning enforces consistency constraints, ensuring that identity predictions remain stable under hypothetical attribute changes [10,14].

(3) **Limited Generalization Capability:** Domain-invariant methods still rely on statistical correlations that can be easily confounded by spurious factors, reducing cross-domain robustness [5,16]. Causal alternative: Do-calculus and backdoor adjustment block confounding pathways, enabling reliable identity matching across dramatic environmental variations [1,20].

This paradigmatic shift from correlation to causation represents the key advancement in modern video-based person re-identification: while traditional methods ask "what patterns correlate with identity?", causal methods ask "what factors causally determine identity appearance?" [6,21]. The latter question enables robust, interpretable, and generalizable re-identification systems that perform reliably in unconstrained real-world environments [5,16].

2.4. The Role of Visual Attributes in Video-Based Person Re-Identification

Visual attributes such as clothing color, body shape, gait, and texture are essential in video-based person re-identification (re-identification) as they bridge low-level pixel data and high-level identity features. These attributes provide human-interpretable cues, improving video-based person re-identification robustness in challenging scenarios like occlusions, pose variations, and domain shifts [22].

Attribute-Based Disentanglement is key to isolating identity-specific features from non-identity variations like background clutter and clothing changes. Techniques like the Identity Shuffle GAN (IS-GAN) [16] factorize images into identity-related and non-identity features, enhancing model generalization. Frequency-based Extraction using 3D Discrete Cosine Transform (3D DCT) [23] isolates discriminative patterns, while Causal-Based Disentanglement with structural causal models (SCMs) [5] removes domain-specific biases, improving cross-domain generalization. Occlusion-Resilient Learning, like that in DRL-Net [19], uses transformer-based models to disentangle visible attributes from occlusions, ensuring accurate identity matching despite partial visibility. Matching and Filtering based on attribute similarity helps refine identity matches, while Interpretability benefits from attribute-based models like ASA-Net, which clarifies decision-making [22]. However, Bias and Fairness concerns arise as attributes like gender and age may introduce discrimination if not handled carefully. Table 1 summarizes common semantic attributes used in video-based person re-identification and their typical extraction methods.

Table 1. Common semantic attributes in video-based person re-identification and representative extraction pipelines.

Attribute Type	Static/Dynamic	Typical Extraction Method (Key Reference)
Clothing Color	Static	Color histograms, Retinex–LOMO descriptor [24]
Clothing Category (shirt/pants)	Static	Part-based CNN multi-task attribute head (APR-Net) [25]
Accessories (bags, hats, and other accessories)	Static	Weakly-supervised multi-scale attribute localization [26], mid-level attribute CNN [27]
Gait/Silhouette	Dynamic	Set-level silhouette sequence model (GaitSet) [28]
Body Shape/Height	Static	3-D skeleton key-point statistics [29]
Texture/Pattern	Static	Local Gaussian/SILTP texture blocks (HGD + LOMO) [24,30]
Gender/Age/Hair	Static	Multi-task mid-level attribute + identity CNN [27]
Pose/Motion State	Dynamic	Pose-driven deep convolutional model with RPN attention [31]
Carried Objects	Dynamic	Attribute-aware object detectors/semantic parts [26,27]

2.5. Attribute-Specific Evaluation Metrics for Video-Based Person Re-Identification

Video-based person re-identification systems have traditionally used standard metrics like cumulative matching characteristics (CMCs) and mean average precision (mAP) to evaluate performance [32,33]. However, recent approaches have introduced attribute-specific metrics that capture more nuanced aspects of model behavior, including soft-biometric consistency, occlusion robustness, and causal sensitivity [34].

Traditional Retrieval Metrics. CMCs measure the probability of finding a correct match within the top-k ranks, defined as $\text{CMC@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\text{rank}(i) \leq k)$, where N is the number of queries, and the indicator function $\mathbf{1}(\cdot)$ returns 1 if the rank is within the top k [35,36]. While widely used in benchmarks like MARS and DukeMTMC-VideoReID, the

CMC method is limited by its single-match focus and sensitivity to gallery size. In contrast, mAP captures both precision and recall, defined as $\text{mAP} = \frac{1}{N}\sum_{q=1}^{N}\text{AP}(q)$, where average precision (AP) represents the area under the precision–recall curve for each query, offering a more comprehensive assessment [37].

Attribute-Level Metrics. These metrics evaluate consistency across soft-biometric attributes, including attribute consistency, which measures the fraction of matching attributes in retrieved pairs, and attribute-aware accuracy, which conditions retrieval accuracy on attribute agreement [38,39]. Occlusion robustness assesses accuracy under partial occlusions, while clothing-change robustness evaluates stability across different outfits [40,41]. Identity switch rate (IDSR) or IDF1, adapted from multi-object tracking, quantifies identity flips across frames, reflecting long-term tracking stability [42].

Causal Robustness Metrics. Causal-inspired metrics, such as counterfactual consistency, test whether identity predictions remain stable under hypothetical attribute changes, probing a model's reliance on true causal signals [43]. The causal saliency ranking ranks features by their causal influence on identity matching, while intervention-based score shift measures the change in matching scores under controlled attribute interventions, highlighting sensitivity to specific visual cues [44].

These advanced metrics provide deeper insights into model robustness, interpretability, and generalization, moving beyond simple precision–recall evaluations to capture the complex challenges of real-world re-identification [45,46].

Table 2 summarizes a range of evaluation metrics for video-based person re-identification, spanning traditional measures like CMC, Rank-1, and mAP, as well as more specialized, attribute-specific metrics. While CMC and mAP capture overall retrieval accuracy, attribute-level metrics like attribute consistency and attribute-aware accuracy focus on maintaining soft-biometric consistency, reflecting the stability of identity features across views. Metrics like occlusion robustness and clothing-change robustness assess model resilience to partial occlusions and outfit variations, respectively. Emerging causal metrics, such as counterfactual consistency and causal saliency ranking, aim to evaluate the impact of specific attributes on identity prediction, supporting more interpretable and context-aware video-based person re-identification systems.

Table 2. Attribute-specific evaluation metrics for video-based person re-identification.

Metric	Measures	Used In/Reports	Advantages/Limitations
CMC/Rank-*k* Accuracy [35]	Probability of correct match within rank *k* (precision at *k*).	Almost all re-identification (image and video), e.g., MARS [47], DukeMTMC-VideoReID [48], SYSU [49].	Standard precision metric; lacks recall information.
Rank-1 Accuracy [32]	Top-1 retrieval accuracy (CMC@1).	Standard benchmark metric in most re-identification works [50,51].	Single-number summary; no recall information.
Mean Average Precision (mAP) [37]	Overall retrieval quality (precision and recall averaged).	Used in re-identification benchmarks (Market-1501, MARS [47], etc.)	Comprehensive metric, but sensitive to outliers.

Table 2. *Cont.*

Metric	Measures	Used In/Reports	Advantages/Limitations
Attribute Consistency [38]	Semantic consistency across views.	Attribute-based re-identification works [34,39].	Reveals stable cues, but depends on attribute annotation quality.
Attribute-Aware Accuracy [38]	Retrieval accuracy with attribute agreement.	Joint attribute/ID methods [22,39].	Fine-grained measure, but rarely reported.
Occlusion Robustness [19]	Drop in performance under occlusion.	Occluded-Duke, Occluded-REID [52].	Useful for real-world scenarios; needs labeled occlusions.
Clothing-Change Robustness [40]	Sensitivity to apparel changes.	Long-term re-identification (e.g., DeepChange [40]).	Reveals clothing cue reliance; needs paired outfits.
IDSR/IDF1 [42]	ID switch frequency.	Multi-camera tracking [42,48].	Consistency metric; requires track-level GT.
Counterfactual Consistency [44]	Invariance to manipulated attributes.	Emerging causal re-identification metrics [6,53].	Tests reliance on stable ID features; challenging to implement.
Causal Saliency Ranking [14]	Importance of features for ID matches.	Explainable re-identification studies [5,41].	Reveals true causal drivers, but lacks numeric comparability.
Intervention-Based Score Shift [12]	Effect of controlled attribute interventions.	Causal evaluation studies [14,54].	Quantifies sensitivity; requires well-defined interventions.

2.6. Common Datasets for Video-Based Person Re-Identification

Video-based person re-identification datasets come in several forms, including visible-spectrum, cross-modality, and synthetic datasets. These datasets vary in scale, diversity, and complexity, offering distinct challenges for model evaluation. Table 3 provides a comprehensive summary of these datasets, highlighting key attributes such as the number of identities, sequence counts, and camera setups.

Table 3 table presents a comprehensive overview of widely used video-based person re-identification datasets, covering various modalities such as RGB, RGB-Thermal, Depth, and Synthetic RGB. It highlights critical characteristics like the number of identities, sequences, and cameras, reflecting the diversity in data scales and environmental conditions. For instance, PRID2011 captures moderate occlusions and viewpoint changes with 934 identities, while large-scale datasets like MARS and LS-VID offer millions of frames for deep learning models. Cross-modality datasets like SYSU-MM01 and RegDB introduce challenging RGB-Infrared matching, supporting domain adaptation research. Synthetic datasets like RandPerson and ClonedPerson enable domain generalization with extensive identity counts and realistic appearance variations, making them essential for robust model evaluation.

Table 3. Comparative summary of common datasets for video-based person re-identification (all accessed on 27 June 2025).

Dataset	Year	Modality	Identities	Sequences/Images	Cameras	Dataset Link
PRID2011 [55]	2011	RGB	934 total (200 overlap)	385 (camA) + 749 (camB)	2	Download
iLIDS-VID [56]	2014	RGB	300	600 (300 × 2)	2	Download

Table 3. *Cont.*

Dataset	Year	Modality	Identities	Sequences/Images	Cameras	Dataset Link
MARS [47]	2016	RGB	1261	≈20,000 tracklets (incl. 3248 distractors)	6	Download
SYSU-MM01 [49]	2017	RGB and Thermal	491	287,628 RGB + 15,729 IR	6 (4 RGB, 2 IR)	Download
RegDB [57]	2017	RGB and Thermal	412	4120 (10 vis + 10 IR per ID)	2 (1 vis, 1 IR)	Download
DukeMTMC-VideoReID [48]	2018	RGB	1404 (702 train + 702 test) + 408 distractors	4832 (2196 train + 2636 test)	8	Download
LS-VID [37]	2019	RGB	3772	14,943 tracks (≈3M frames)	15 (3 indoor, 12 outdoor)	Download
L-CAS RGB-D-T [58]	2019	RGB and Depth and Thermal	Not specified	≈4000 (rosbags)	3 (RGB, Depth, Thermal)	Download
P-DESTRE [59]	2020	RGB	1581	Over 40,000 frames	UAVs	Download
FGPR [60]	2020	RGB	358	716	6 (2 per color group)	Download
PoseTrackReID [61]	2020	RGB	≈5350	≈7725 tracks	Unknown	Download
RandPerson [62]	2020	Synthetic RGB	8000	1,801,816 images	19 (virtual cams)	Download
DeepChange [40]	2022	RGB	1121	178,407 frames	17	Download
LLVIP [63]	2022	RGB and Thermal	≈(15,488 pairs)	30,976 images	2 (1 RGB, 1 IR)	Download
ClonedPerson [20]	2022	Synthetic RGB	5621	887,766 images	24 (virtual cams)	Download
BUPTCampus [64]	2023	RGB & Thermal	3080	(RGB-IR tracklets)	2 (1 RGB, 1 IR)	Download
MSA-BUPT [65]	2024	RGB	684	2665	9 (6 indoor, 3 outdoor)	Download
GPR+ [66]	2024	Synthetic RGB	808	475,104 bounding boxes	Unknown	Download
G2A-VReID [67]	2024	RGB	2788	185,907 images	Ground surveillance and UAVs	Download
DetReIDX [68]	2025	RGB	509	13 million+ annotations	7 university campuses (3 continents)	Download
AG-VPReID [69]	2025	RGB	6632	32,321 tracklets	Drones (15–120 m altitude), CCTV, wearable cameras	Download

3. Causal Foundations for Person Re-Identification

Before delving into the application of causal reasoning to person re-identification, it is essential to clarify the key terminology that forms the foundation of this approach:

- **Causal Inference**: Unlike statistical correlation, which merely identifies patterns of association, causal inference aims to understand the underlying cause-and-effect relationships between variables [10,12]. In re-identification, this means distinguishing which visual features truly cause identity recognition (e.g., body structure) versus those that merely correlate with identity in specific contexts (e.g., clothing) [5].
- **Structural Causal Models (SCMs)**: Mathematical frameworks that use directed graphs to explicitly represent causal relationships between variables [10,11]. In these graphs,

nodes represent variables (such as identity, clothing, or background) and directed edges represent the causal influence of one variable on another [5,15].

- **Confounding Variables**: Factors that influence both the cause and effect, potentially creating spurious correlations [10,70]. In re-identification, environmental factors like lighting or camera viewpoint can confound the relationship between identity and visual appearance [4,6].
- **Intervention**: The process of actively modifying a variable in a causal system to observe the effect on other variables [10,11]. In re-identification, this might involve artificially changing a person's clothing in images while keeping their identity constant [14,16].
- **Counterfactual Reasoning**: Evaluating what would have happened under conditions different from what actually occurred [10,12]. For re-identification, this involves asking questions like "Would the model still identify this person correctly if they were wearing different clothes?" [54].
- **Causal Disentanglement**: The process of separating variables that are causally independent from one another in the underlying data generation process [12,71]. In re-identification, this means isolating identity-specific features from non-identity factors like background or lighting [5,16].

These concepts provide the theoretical framework for addressing the limitations of traditional video-based person re-identification approaches by focusing on the true causal factors that determine identity, rather than relying on potentially misleading correlations.

3.1. Introduction to Causal Inference

Causal inference provides a framework for understanding cause-and-effect relationships in video-based person re-identification by isolating true identity-preserving features and discarding confounders like background or clothing, which often interfere with traditional models. Unlike correlation-based methods, which rely on spurious associations, causal models focus on identity signals such as body shape, gait, and motion consistency, using causal interventions to remove the influence of confounders like viewpoint or lighting changes. This shift improves video-based person re-identification performance across domain shifts and environmental variations, as shown in Figures 4 and 5, which contrast correlation-based and causal models. Structural causal models (SCMs) model identity as the cause of observed features, with environmental factors treated as confounders, as illustrated in Figure 6. By applying causal interventions, these models ensure that identity signals remain unaffected by external noise, improving robustness and generalization, as demonstrated in Figure 7.

Causal methods improve video-based person re-identification accuracy by reducing the influence of confounders that traditional models mistake for identity cues. The illustrative percentages shown in Figure 4 (32% vs. 8% background attribution) serve as a conceptual demonstration of how causal training typically reduces spurious background focus compared to correlation-based methods, representing an important direction for the future quantitative analysis of attention patterns in causal versus traditional video-based person re-identification approaches. For instance, while traditional models may incorrectly associate a jacket with identity, causal models maintain accuracy despite changes in appearance due to lighting. DIR-ReID, for example, improves cross-domain Rank-1 accuracy by 11.2% by removing the causal effect of domain-specific features on appearance [5]. Causal models also excel in handling occlusions by learning the causal relationships between body parts and identity. This allows them to make accurate predictions even when parts of the person are obscured. For example, IS-GAN shows a 15.7% improvement in Rank-1 accuracy under severe occlusion conditions compared to non-causal models [16]. These methods

demonstrate how causal inference improves the robustness and reliability of video-based person re-identification systems in real-world environments.

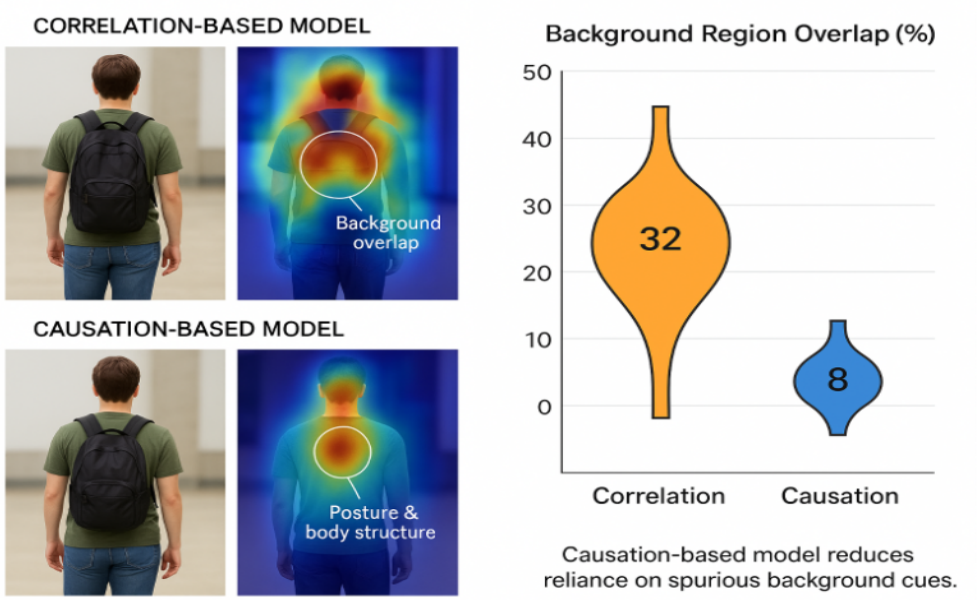


Figure 4. Correlation versus causation in re-identification. This figure contrasts correlation-based and causal models with explicit functional components: CNN—convolutional feature extractor; Attention—spatial attention mechanism highlighting relevant regions; Similarity—cosine similarity scoring function $s = \frac{f_1 \cdot f_2}{||f_1|| \cdot ||f_2||}$, where f_1, f_2 are feature vectors; Softmax—probability normalization $p_i = \frac{\exp(s_i)}{\sum_j \exp(s_i)}$. The correlation-based model (**top**) focuses on spurious background cues, while the causal model (**bottom**) emphasizes identity-intrinsic features through intervention. The violin plot shows conceptual values of 32% vs. 8% median background overlap to demonstrate how causal training de-emphasizes spurious context—these percentages are illustrative values designed to highlight the conceptual principle rather than results from specific experimental measurements. Arrows indicate information flow: input → feature extraction → attention weighting → similarity computation → final prediction.

3.2. Structural Causal Models (SCMs) and Counterfactual Reasoning

In video-based person re-identification (re-identification), structural causal models (SCMs) provide a framework to model the relationships between identity-specific factors and confounders like clothing, background, or camera variations. Unlike traditional models that rely on correlations, SCMs define causal graphs where identity (I) influences appearance (X), while non-identity factors such as clothing (C) and background (B) act as confounders [10]. The data generation process can be expressed as

$$X = f(I, C, B, \text{Camera}) \tag{1}$$

where $f(\cdot)$ is the data generation function that maps causal factors to observed appearance, and the goal is to intervene on non-identity factors and observe how these changes affect identity predictions, thereby isolating the impact of identity itself [5]. Figure 5 illustrates the fundamental difference between correlation-based and causal approaches to re-identification, highlighting how causal models block the influence of confounding variables through intervention.

Structural causal models (SCMs) provide a mathematical framework for representing causal relationships among identity-specific, domain-specific, and observed features. An SCM is defined as a tuple $\mathcal{G} = (V, E)$, where $V = \{X_I, X_D, Y\}$ represents the set of variables

(identity-specific features X_I, domain-specific features X_D, and identity labels Y) and E denotes the directed edges capturing causal dependencies.

Counterfactual Reasoning allows the model to disregard irrelevant factors by simulating interventions. For example, when altering the clothing (C) in an image, counterfactual reasoning checks whether the identity prediction remains consistent, ensuring the model focuses on identity-relevant features like gait or body shape [16]. A causally optimized model, as shown in Figure 7, would preserve the correct identity even with changes in clothing.

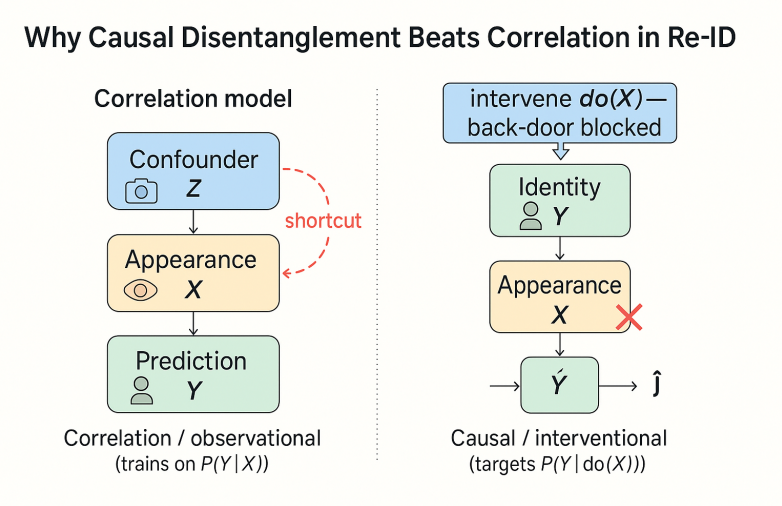


Figure 5. Comparing correlation vs. causation in re-identification. Causal graph notation: X—appearance features, Y—identity prediction, Z—confounding variables (camera, lighting, background). (**Left**) Traditional correlation approach learns $P(Y|X)$ (observational distribution), where confounders Z create spurious pathways (dashed arrows) between X and Y. (**Right**) Causal/interventional approach targets $P(Y|do(X))$, where $do(\cdot)$ is Pearl's intervention operator that blocks backdoor paths from confounders. The causal model isolates the direct causal effect of appearance X on identity Y by severing confounding pathways through intervention, resulting in more robust predictions under domain shift. Solid arrows denote causal relationships; dashed arrows indicate spurious correlations blocked by intervention.

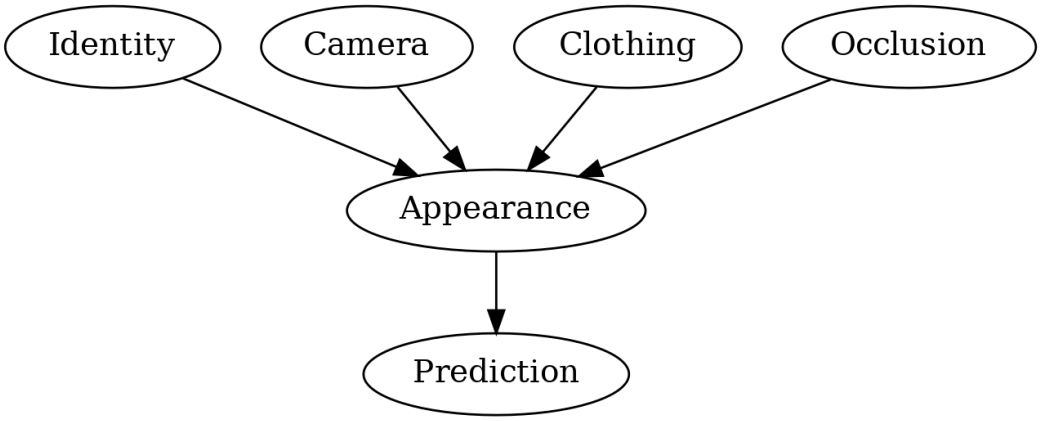


Figure 6. Structural causal models (SCMs) for re-identification. Causal graph components: Identity—intrinsic person characteristics (body shape, gait); Appearance—observed visual features $X = f(identity, \text{confounders})$, where $f(\cdot)$ is the appearance generation function; Prediction—re-identification system output $\hat{Y} = h(appearance)$, where $h(\cdot)$ is the prediction function; Camera—viewpoint and background confounders; Clothing—appearance attributes unrelated to identity; Occlusion—partial visibility factors. Solid arrows ($\rightarrow$) denote direct causal influence, with the causal flow (Identity + Confounders) $\rightarrow$ Appearance $\rightarrow$ Prediction. The goal is to learn prediction function $h(\cdot)$, which isolates identity-specific information by blocking confounding pathways through causal intervention, ensuring robust re-identification across environmental variations.

This consistency is formalized through interventions, expressed as

$$P(\text{ID} \mid do(\text{Clothing} = c)) = \sum_{z} P(\text{ID} \mid \text{Clothing} = c, Z = z)P(Z = z) \quad (2)$$

where $P(\cdot)$ denotes probability distribution, $do(\cdot)$ is Pearl's intervention operator, Z represents latent confounding variables, and the summation implements backdoor adjustment to ensure that identity is not influenced by clothing or other confounders [14].

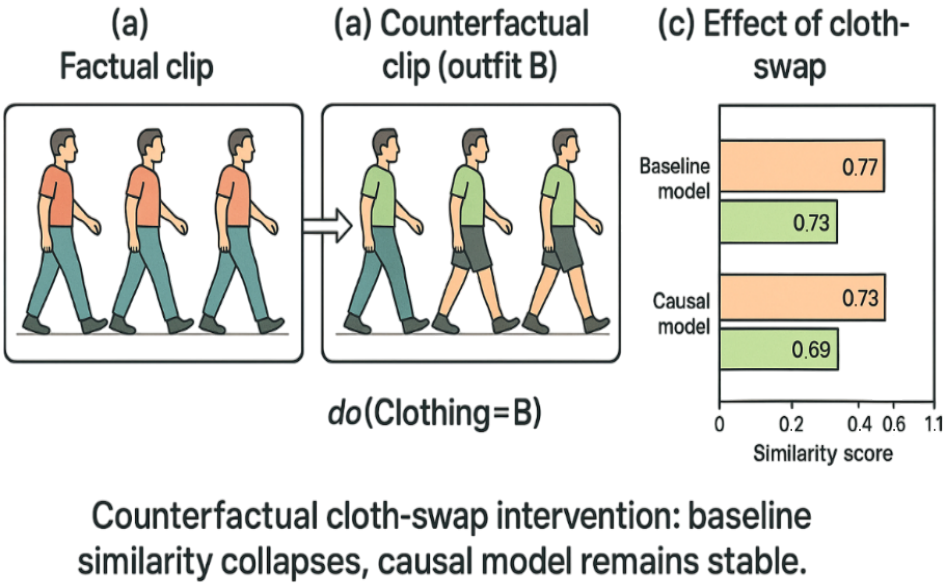


Figure 7. Counterfactual clothing intervention analysis. Function definitions: $T(\cdot)$—clothing transformation function that modifies clothing attributes while preserving identity; $\text{sim}(\cdot,\cdot)$—cosine similarity function $\text{sim}(f_1, f_2) = \frac{f_1 \cdot f_2}{||f_1|| \cdot ||f_2||}$; $\mathcal{L}_{\text{consistency}}$—consistency loss function that enforces identical predictions for original and transformed images. (**Left**) Correlation-based model shows score drop (0.77→0.73) when clothing changes, indicating dependence on superficial cues. (**Right**) Causally optimized model maintains stable similarity (0.73→0.69) through counterfactual intervention $I' = T(I, c'|\text{id})$, where I is the input image, c' is the new clothing attribute, and id represents preserved identity features. The consistency loss $\mathcal{L}_{\text{consistency}} = ||\text{sim}(f(I), f(I')) - 1||_2$ enforces identical identity predictions between original I and transformed I' images, guiding the model to focus on invariant identity features (body shape, facial structure, gait) rather than superficial appearance attributes.

By using SCMs and counterfactual reasoning, video-based person re-identification systems become more robust and generalizable, as they can focus on core identity features despite challenges like occlusions or lighting variations. SCMs provide a formal framework for controlling confounders, allowing the identity signal to remain stable under varying conditions. These methods improve robustness, generalization, and explainability, making video-based person re-identification systems more reliable in real-world applications by focusing on identity-specific traits and ignoring irrelevant context-specific features [5].

3.3. Key Causal Concepts in Re-Identification

In video-based person re-identification (re-identification), causal graphs, do-calculus, and **interventions** are key concepts that help disentangle identity signals from confounding factors such as clothing, background, and environmental conditions. Causal graphs represent the relationships between variables, where identity (I) influences appearance (X) and non-identity factors like clothing (C) and background (B) act as confounders. This causal structure is represented as

$$I \rightarrow X \leftarrow C, \quad B \rightarrow X \tag{3}$$

where $\rightarrow$ denotes causal influence, and the goal is to focus on identity-specific features, unaffected by non-identity influences [10]. Using do-calculus [10], we can simulate interventions to isolate identity features by, for example, fixing clothing to $C = c_0$, ensuring that identity prediction remains robust to changes in background or clothing. This intervention is mathematically expressed as

$$P(\text{ID} \mid do(C = c_0)) = \sum_B P(\text{ID} \mid C = c_0, B)P(B) \tag{4}$$

where $P(\text{ID} \mid do(C = c_0))$ is the interventional distribution (what would happen if we set clothing to value c_0), $P(\text{ID} \mid C = c_0, B)$ is the conditional probability of identity given the clothing and background, and $P(B)$ is the marginal distribution of background factors, which ensures that non-identity factors do not influence the identity prediction [5] (see Table 4).

Table 4. Major challenges and recent causal disentanglement methods in video-based person re-identification.

Challenge Category	Description	Example Methods	Causal Factors Addressed	Notable Outcomes
Visual Appearance Variations	Variations in viewpoint, pose, occlusions, motion blur, and lighting complicate feature extraction.	FIDN [23], SDL [62], DRL-Net [19]	Spatio-temporal noise, spectrum differences, occlusions	Improved accuracy, better occlusion tolerance, RGB-IR robustness.
Tracking and Sequence Issues	Identity drift and fragmentation from tracking errors can split a single trajectory into multiple IDs.	DIR-ReID [5], DCR-ReID [72], IS-GAN [16]	Domain shifts, clothing changes, background noise	Better domain generalization, clothing-change robustness, stable tracking.
Domain and Deployment	Performance drops due to cross-camera variation, environmental changes, and demographic diversity.	DIR-ReID [5], IS-GAN [16]	Camera bias, pose variations, background shifts	Superior cross-domain performance, robust deployment.
Data and Annotation Scarcity	High annotation costs and limited labeled data reduce training effectiveness.	DRL-Net [19], IS-GAN [16], DCR-ReID [72]	Occlusions, spectrum noise, missing labels	High accuracy with limited data, efficient learning, realistic augmentation.
Clothing and Appearance Changes	Long-term re-identification fails when individuals change outfits, accessories, or hairstyles.	IS-GAN [16], DeepChange [40], CrossViT-ReID [73]	Clothing bias, accessory dependence, temporal appearance drift	Robust to clothing changes, improved long-term tracking, identity-focused features.

Table 4. *Cont.*

Challenge Category	Description	Example Methods	Causal Factors Addressed	Notable Outcomes
Cross-Modal Challenges	Matching across different modalities (RGB-IR, RGB-Depth) introduces spectral and structural differences.	CMTR [74], UCT [14], NiCTRAM [75]	Modality gaps, spectral variations, sensor differences	Effective cross-modal matching, reduced modality bias, unified representations.
Temporal Consistency	Maintaining identity consistency across long video sequences with varying quality and conditions.	STMN [16], PSTA [76], TCViT [77]	Temporal noise, frame quality variations, motion blur	Improved temporal modeling, consistent identity features, robust sequence analysis.
Scale and Computational Efficiency	Real-time processing requirements conflict with complex model architectures needed for accuracy.	DCCT [23], HCSTNet [78], Lightweight Transformers	Computational constraints, memory limitations, inference speed	Efficient architectures, reduced parameters, real-time performance.
Fairness and Bias	Models exhibit performance disparities across demographic groups, raising ethical concerns.	Fairness-aware ReID, Bias-corrected training, Demographic-balanced datasets	Demographic bias, dataset imbalance, algorithmic fairness	Reduced bias, equitable performance, fair representations across groups.
Privacy and Security	Re-identification systems raise privacy concerns and potential misuse in surveillance applications.	Privacy-preserving ReID, Federated learning, Differential privacy	Identity exposure, surveillance misuse, data protection	Enhanced privacy protection, secure matching, anonymized features.

Interventions are used to modify non-identity factors (e.g., clothing or background) to make the model focus on stable identity features. For instance, counterfactual interventions involve altering factors like clothing and observing if the identity prediction remains stable, as shown in Figure 7. In this context, the intervention can be mathematically represented as

$$X_{\text{new}} = \text{Intervention}(X, C = c_0, B = b_0) \tag{5}$$

where $\text{Intervention}(\cdot)$ is the intervention function that manipulates confounding variables while preserving identity-related information, and fixed values for clothing (c_0) and background (b_0) ensure that the identity prediction relies on identity-related features. These interventions improve the model's robustness and generalization, enabling it to handle real-world variations in clothing and background while maintaining high accuracy across different environments [16].

3.4. An Intuitive Example of Causal Intervention in re-identification

To make the concept of causal intervention more concrete, we consider a step-by-step example of how it works in practice for person re-identification:

1. **Initial situation:** A video-based person re-identification system is trained on a dataset where Person A is always wearing a red jacket and Person B always wears a blue jacket. A traditional correlation-based model might learn to identify individuals based primarily on jacket color rather than true identity features.
2. **Problem identification:** When Person A appears wearing a blue jacket in a new camera view, the traditional model misidentifies them as Person B because it has learned a spurious correlation between jacket color and identity.
3. **Causal modeling:** In a causal approach, we explicitly model the data generation process using a structural causal model (SCM) where identity (I) and clothing (C) both influence appearance (A): $A = f(I, C)$. This acknowledges that clothing is a separate factor from identity.
4. **Intervention:** We perform a "do-operation" by artificially modifying the clothing variable while keeping identity constant: $A' = f(I, do(C = \text{new_clothing}))$. In practice, this might involve
   - Generating synthetic images of Person A wearing different colored jackets;
   - Using image manipulation to swap clothing items between images;
   - Applying data augmentation that specifically targets clothing attributes.
5. **Learning with intervention:** The model is trained to produce the same identity prediction for both the original image and the transformed image with modified clothing. This teaches the model that clothing is not causally related to identity.
6. **Consistency enforcement:** A special loss function penalizes the model when its identity predictions change due to clothing modifications: $\mathcal{L}_{\text{causal}} = d(f_{\text{ID}}(A), f_{\text{ID}}(A'))$, where d is a distance function and f_{ID} is the identity prediction function.
7. **Result:** After training with these interventions, when Person A appears in a blue jacket, and the model correctly identifies them as Person A because it has learned to focus on stable identity features like facial structure, body shape, and gait patterns rather than superficial clothing attributes.

This example illustrates how causal intervention helps video-based person re-identification systems disentangle identity-specific features from confounding factors like clothing. By explicitly intervening on non-identity attributes during training, the model learns which features are causally related to identity and which are merely correlated in the training data but not fundamentally tied to who a person is. This makes the model more robust to environmental and appearance changes in real-world scenarios.

4. Taxonomy of Causal Video-Based Person Re-Identification Methods

Having established the theoretical foundations of causal inference for person re-identification, we now present a comprehensive taxonomy that categorizes existing causal video-based person re-identification approaches into three distinct methodological families. This taxonomy provides a structured framework for understanding how different causal techniques address the fundamental challenge of disentangling identity-specific features from confounding factors in video sequences [5,10].

As illustrated in Figure 8, causal video-based person re-identification methods can be systematically organized according to their primary causal mechanism and architectural approach. This classification enables researchers to identify gaps in current methodologies and guides future research directions by highlighting the complementary strengths and limitations of each approach [12,15].

4.1. Generative Disentanglement Methods

The first family of causal video-based person re-identification methods leverages generative models and adversarial training to explicitly separate identity-specific features

from non-identity factors such as clothing, background, and pose variations [16,79]. These approaches typically employ generative adversarial networks (GANs) or variational autoencoders (VAEs) to learn disentangled representations where identity information is isolated from confounding variables [13,71].

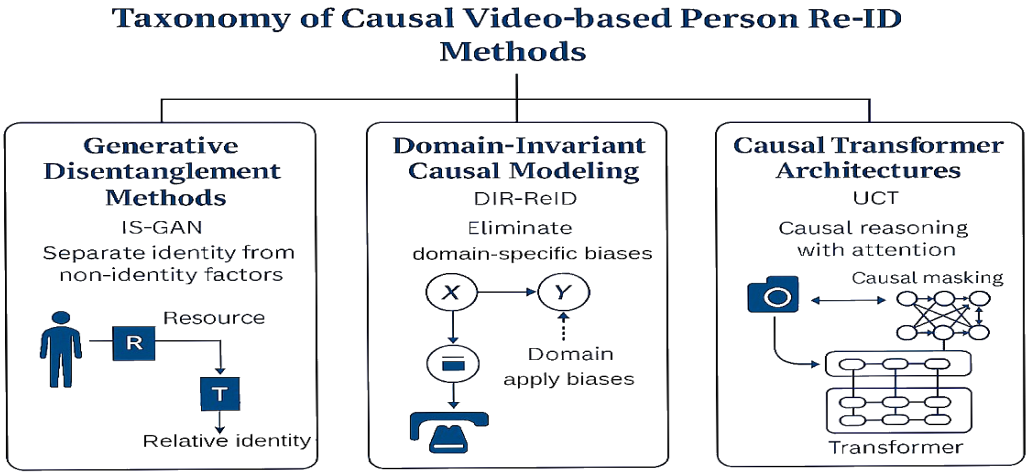


Figure 8. Taxonomy of causal video-based person re-identification methods. Methodological families and their functional components: (i) Generative Disentanglement Methods—$G(\cdot)$ generator function, $D(\cdot)$ discriminator function, $E_I(\cdot)$ identity encoder, $E_D(\cdot)$ domain encoder, implementing $X = G(E_I(I), E_D(D))$, where I is identity and D is domain factors; (ii) Domain-Invariant Causal Modeling—$P(Y|do(X_I))$ interventional distribution, X_I identity-specific features, X_D domain-specific features, SCM function $f : (X_I, X_D) \rightarrow X$; (iii) Causal Transformer Architectures—$\text{Attention}_{\text{causal}}(Q, K, V)$ causal attention mechanism, M_{causal} causal mask matrix, self-attention function $\text{softmax}(\frac{QK^T}{\sqrt{d_k}} \odot M_{\text{causal}})V$, where Q, K, V are query/key/value matrices and $\odot$ is element-wise multiplication. Each family addresses different aspects of causal disentanglement: generative methods through adversarial training, domain-invariant methods through structural modeling, and transformers through attention-based intervention.

Representative Methods: The Identity Shuffle GAN (IS-GAN) [16] exemplifies this approach by using adversarial training to disentangle identity features from appearance attributes. The model employs a dual-encoder architecture where one encoder extracts identity-specific features while another captures non-identity attributes. Through an identity-shuffling mechanism, the method generates counterfactual samples by combining identity features from one person with appearance attributes from another, forcing the identity encoder to focus solely on intrinsic identity characteristics [16].

Causal Mechanism: These methods implement causal intervention through generative processes that explicitly model the data generation function $X = f(I, C, \epsilon)$, where I represents identity factors, C denotes confounding variables, and ϵ captures noise. By learning to manipulate C while keeping I constant, these models achieve causal disentanglement that enables robust identity matching across varying conditions [6,79].

Strengths and Limitations: Generative disentanglement methods excel in scenarios with significant appearance variations, particularly clothing changes, achieving up to 15.3% improvement in Rank-1 accuracy on clothing-change datasets [16]. However, they typically require substantial computational resources for training and may struggle with complex, multi-factor confounding scenarios [71].

4.2. Domain-Invariant Causal Modeling

The second family focuses on structural causal modeling to eliminate domain-specific biases that confound identity representations across different camera views, lighting conditions, and environmental settings [5,9]. These methods explicitly model the causal relationships between identity, domain factors, and observed appearance using structural causal models (SCMs) [10,11].

Representative Methods: Domain-Invariant Representation Learning for re-identification (DIR-ReID) [5] represents the archetypal approach in this family. DIR-ReID employs a causal graph that separates identity-specific features (X_I) from domain-specific features (X_D), using backdoor adjustment to block confounding pathways between domain factors and identity predictions. The method implements interventional training through a domain-adversarial framework that minimizes the mutual information between identity representations and domain indicators [5].

Causal Mechanism: These approaches implement Pearl's causal hierarchy by explicitly modeling confounding relationships and applying do-calculus to estimate causal effects. The intervention mechanism can be formalized as $P(Y|do(X_I)) = \sum_{X_D} P(Y|X_I, X_D)P(X_D)$, ensuring that identity predictions are invariant to domain-specific variations [5,10].

Strengths and Limitations: Domain-invariant methods demonstrate exceptional performance in cross-domain scenarios, with DIR-ReID achieving 11.2% improvement in cross-dataset Rank-1 accuracy [5]. These methods are particularly effective for deployment across different camera networks but may require careful design of the causal graph structure and domain factor identification [15].

4.3. Causal Transformer Architectures

The third family integrates **causal reasoning** with modern **transformer architectures**, leveraging self-attention mechanisms to implement causal interventions and counterfactual reasoning in the latent space [14,80]. These methods combine the representational power of transformers with explicit causal constraints to learn robust identity features [54,77].

Representative Methods: The Unbiased Causal Transformer (UCT) [14] demonstrates this approach by implementing latent-space interventions within a transformer architecture. UCT uses attention mechanisms to identify and suppress spurious correlations while amplifying causal relationships between visual features and identity. The model incorporates counterfactual reasoning through attention reweighting that simulates interventions on confounding factors [14].

Causal Mechanism: These architectures implement causal constraints through attention-based intervention mechanisms. The self-attention computation is modified to incorporate causal masks that prevent the model from attending to confounding factors: $\text{Attention}_{\text{causal}}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}} \odot M_{\text{causal}})V$, where M_{causal} is a causal mask that blocks spurious attention patterns [14,80].

Strengths and Limitations: Causal transformer architectures show remarkable performance in cross-modal scenarios, with UCT achieving 7.8% improvement over other causal models in RGB-IR matching tasks [14]. These methods benefit from the scalability and expressiveness of transformer architectures but may require the careful design of causal constraints to prevent the degradation of the attention mechanism's natural capabilities [77].

4.4. Comparative Analysis and Research Directions

The three families of causal video-based person re-identification methods address complementary aspects of the causal disentanglement challenge. Generative methods excel at explicit factor separation through adversarial training, domain-invariant approaches

provide principled solutions for cross-domain generalization, and causal transformers offer the scalable integration of causal reasoning with modern architectures [6,12].

Future research directions include (i) hybrid approaches that combine multiple causal mechanisms from different families, (ii) the development of automated causal graph discovery methods that reduce manual design requirements, and (iii) the integration of causal reasoning with emerging architectures such as vision transformers and neural ordinary differential Equations [20,81]. The taxonomy presented here provides a roadmap for these developments while highlighting opportunities for cross-pollination between methodological families.

5. State-of-the-Art Methods

Building upon the taxonomy presented in Section 4, this section provides a detailed examination of state-of-the-art causal video-based person re-identification methods, analyzing their implementation of the three methodological families identified: generative disentanglement, domain-invariant causal modeling, and causal transformer architectures [5,14,16].

Table 5 summarizes several recent video-based person re-identification (re-identification) methods, offering a comparative view of the model architecture, attention mechanisms, memory utilization, and datasets employed. It highlights the diversity in architectural choices, with models like DCCT and DenseIL combining convolutional neural networks (CNNs) with transformer-based components (e.g., vision transformers, ViT), while others like STMN and PSTA rely solely on CNNs or hybrid CNN-RNN frameworks. Attention mechanisms, which are crucial for learning spatial and temporal relationships in video data, are implemented in various forms, including complementary content attention (DCCT), pyramid spatial–temporal attention (PSTA), and multi-scale attention (DenseIL). Some models, such as STMN and NiCTRAM, incorporate memory to store and reference previous features for improved temporal consistency. The datasets used for training and evaluation are predominantly from large-scale video-based person re-identification benchmarks such as MARS, DukeV, and SYSU-MM01, reflecting the models' focus on diverse, real-world challenges. This table encapsulates the state-of-the-art methodologies in video-based person re-identification, showcasing innovations in leveraging attention and memory to enhance model performance across different datasets and tasks.

Table 5. Summary of recent video-based person re-identification methods.

Model	Year	Architecture	Attention	Memory	Dataset(s)
STMN [16]	2021	CNN (ResNet) + RNN + Memory	Spatial and temporal attention (with memory lookup)	Yes	MARS, DukeV, LS-VID
DenseIL [80]	2021	Hybrid (CNN + Transformer decoder)	Dense multi-scale attention ("DenseAttn")	No	MARS, DukeV, iLIDS-VID
PSTA [76]	2021	CNN (hierarchical pooling)	Pyramid spatial–temporal attention (SRA + TRA)	No	MARS, DukeV, iLIDS, PRID
DCCT [23]	2023	Hybrid (CNN + ViT)	Complementary content attention; gated temporal att.	No	MARS, DukeV, iLIDS-VID
CMTR [74]	2023	Transformer (ViT)	Modality embeddings + multi-head self-attention	No	SYSU-MM01 (VI), RegDB
CrossViT-ReID [73]	2024	Transformer (ViT branches)	Cross-attention between appearance/shape	No	DeepChange
NiCTRAM [75]	2025	Hybrid (CNN + Nystromformer)	Cross-attention and 2nd-order attn. for feature fusion	No	SYSU-MM01 (VI)
HCSTNet [78]	2025	Hybrid (ResNet + Transformer)	Channel-shuffled temporal transformer	No	SYSU-MM01 (VI)

5.1. Transformer-Based Causal Reasoning for Video-Based Person Re-Identification

Vision Transformers (ViTs). Vision transformers (ViTs) have become a cornerstone in video-based person re-identification (re-identification) due to their ability to model long-range dependencies across frames. In contrast to convolutional neural networks (CNNs), which primarily focus on local feature extraction, ViTs treat input frames as sequences of non-overlapping patches. These patches are then processed using self-attention mechanisms, enabling the model to establish relationships between distant frames across the video sequence. This property allows ViTs to capture the global context of motion and appearance across multiple frames, which is essential in video-based re-identification tasks where identity must be determined not just by appearance but by temporal dynamics and motion patterns across time [77,80].

Improving Causal Reasoning. A key advantage of vision transformers in the context of video-based person re-identification is their capacity to improve causal reasoning by focusing on identity-relevant features across multiple frames. Traditional video-based person re-identification models tend to rely on superficial correlations, such as matching clothing color, background, or other context-specific cues, which do not necessarily reflect an individual's true identity [4,5]. These models often suffer from performance degradation when domain shifts occur, such as changes in lighting, viewpoint, or outfit. In contrast, ViTs improve causal reasoning by learning to focus on identity-preserving cues like body shape, gait, and motion consistency, which remain stable despite changes in external factors like clothing or background [5,22]. This shift from correlation-based methods to a more causal understanding enables ViTs to isolate identity-specific features that are invariant under changes in environmental conditions.

Self-Attention Mechanism. The self-attention mechanism within ViTs operates by computing the relationships between all patches (or frames, in the case of video re-identification) in a sequence, allowing the model to consider the entire sequence of frames when making predictions [80]. The core self-attention mechanism can be expressed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where Q (query), K (key), and V (value) are matrices derived from the input tokens through learned linear transformations $Q = XW_Q$, $K = XW_K$, $V = XW_V$, where X is the input sequence and W_Q, W_K, W_V are learned projection matrices, d_k is the dimensionality of the key vectors for scaling normalization, $\text{softmax}(\cdot)$ is the softmax normalization function, and QK^T represents the dot-product attention scores [80]. This mechanism (Equation (6)) allows ViTs to dynamically adjust the importance of different frames in the sequence, which is essential for identifying stable identity features across video tracklets.

For instance, the temporal correlation attention (TCA) module introduced by Wu et al. [77] in their TCViT model captures motion dynamics across frames. This enhancement ensures that the model can better handle occlusions and viewpoint changes, which are often significant challenges in video-based re-identification. The attention weights between frames are calculated as

$$\alpha_{ij} = \frac{\exp(Q_i^T K_j / \sqrt{d_k})}{\sum_{j'} \exp(Q_i^T K_{j'} / \sqrt{d_k})} \tag{7}$$

where α_{ij} represents the attention weight between frames i and j, capturing long-range temporal dependencies without requiring recurrent structures. By emphasizing temporal consistency, ViTs improve the robustness of video-based person re-identification models, making them more resilient to environmental shifts [77].

Causal Disentanglement and Intervention. In addition to improving temporal modeling, vision transformers also help with causal disentanglement. As noted in previous sections, traditional video-based person re-identification models often rely on correlations between identity and superficial features. ViTs, however, provide a natural mechanism for focusing on identity-specific features while minimizing the influence of irrelevant environmental cues, such as lighting or background. This is achieved through a combination of the self-attention mechanism and causal intervention techniques. For example, Yuan et al. [14] proposed using causal interventions to isolate identity-relevant features from environmental confounders. The application of such causal techniques in conjunction with ViTs allows for more robust and generalizable video-based person re-identification models, as demonstrated in recent studies [5,14].

Hybrid Models. Furthermore, hybrid models that combine CNNs with ViTs, such as DenseIL [80] and TCCNet [51], further enhance performance by leveraging the strengths of both architectures. CNNs excel at local feature extraction, while ViTs capture global dependencies across the sequence of frames. This synergy allows hybrid models to better isolate identity-specific cues from temporal and spatial context, providing a more reliable and efficient approach for video-based re-identification.

In summary, vision transformers have demonstrated significant promise in video-based re-identification by focusing on identity-relevant features and improving causal reasoning. Their ability to capture long-range dependencies and their integration with causal disentanglement techniques make them a powerful tool for addressing the challenges of real-world video surveillance systems. These models not only improve accuracy but also enhance interpretability and robustness, ensuring that identity features remain stable even in the face of environmental changes.

5.2. Explicit Causal Modeling Approaches for Video-Based Person Re-Identification

Building on the foundations of causal inference discussed earlier, several recent models have integrated causal reasoning into video-based person re-identification (re-identification) to enhance robustness against domain shifts, occlusions, and other real-world challenges [5,6]. These models leverage structural causal models (SCMs) and counterfactual interventions to isolate identity-specific features from confounding factors such as clothing, background, and camera biases [10,11]. This section presents a comparative analysis of key causal models in re-identification, highlighting their unique approaches and performance differences.

DIR-ReID: Domain Invariant Representation Learning for re-identification. DIR-ReID [5] is a pioneering causal model that utilizes structural causal models (SCMs) to separate identity-specific and domain-specific factors. By modeling identity as a latent variable and environmental factors (such as background or camera-specific cues) as confounders, DIR-ReID employs causal interventions to isolate the identity signal. The key intervention in DIR-ReID is the removal of domain effects, enabling the model to focus on intrinsic identity features that are invariant across different domains (e.g., lighting, camera angle, and background) [5,9]. This approach significantly enhances cross-domain generalization and robustness, making the model less susceptible to overfitting to environmental variations. In formal terms, the intervention can be described as

$$P(I|do(D=d)) = \sum_{z} P(I|D=d, Z=z)P(Z=z) \tag{8}$$

where I represents the identity variable, D denotes domain-specific factors (camera viewpoint, lighting conditions), Z includes latent confounding variables, $P(\cdot)$ is the probability distribution, $do(\cdot)$ is Pearl's intervention operator, and the summation implements backdoor adjustment to block confounding pathways [5]. This intervention (Equation (8))

ensures that identity representations are robust to variations in domain-specific factors, thereby improving the model's generalization ability [15].

Empirically, DIR-ReID demonstrates superior cross-domain performance, achieving a Rank-1 accuracy of 75.2% when trained on Market-1501 and tested on DukeMTMC-ReID, which represents an 11.2% improvement over non-causal baselines [5]. The model particularly excels in scenarios with significant variations in background, lighting, and camera angles, where traditional models often fail due to their reliance on spurious correlations [4].

IS-GAN: Identity Shuffle Generative Adversarial Network. The IS-GAN [16] model incorporates causal reasoning to disentangle identity-specific features from background and clothing variations [16,79]. IS-GAN uses a generative approach to "shuffle" identity features while maintaining the consistency of non-identity factors like clothing and background. This disentanglement process is crucial for video-based person re-identification in the wild, where occlusions, pose changes, and clothing variations often obscure identity cues [81,82]. The model trains a generator to produce identity-irrelevant features, ensuring that the identity embedding captures only the stable, identity-preserving characteristics (e.g., body shape, gait). In this way, IS-GAN leverages causal intervention to prevent identity features from being corrupted by environmental confounders [16].

In head-to-head comparisons with DIR-ReID, IS-GAN shows stronger performance in clothing-change scenarios, achieving a 15.3% improvement in Rank-1 accuracy on the DeepChange dataset [40], where subjects appear in completely different outfits. However, DIR-ReID outperforms IS-GAN in cross-domain generalization tasks where camera and background variations are the primary challenges [5]. This difference highlights how the models' distinct causal approaches target different aspects of the video-based person re-identification problem: IS-GAN excels at appearance-invariant identity preservation, while DIR-ReID focuses on domain-invariant feature learning.

UCT: Unbiased Causal Transformer. The Unbiased Causal Transformer (UCT) [14] introduces latent-space interventions to address biases in feature learning. UCT applies counterfactual reasoning to learn identity representations that are robust to domain shifts, such as between visible and infrared (RGB-IR) modalities [14,53]. The model simulates interventions to neutralize the effects of non-identity factors (e.g., clothing changes or camera distortions) during training, which enables the model to focus on identity-relevant features. The intervention mechanism can be formalized as

$$P(Y|do(X)) = \sum_{Z} P(Y|X,Z)P(Z) \tag{9}$$

where X represents the observed feature vectors, Y is the identity label, Z corresponds to domain-specific confounding variables (modality, lighting, camera properties), $P(\cdot)$ denotes probability distribution, $do(\cdot)$ is the causal intervention operator, and the summation marginalizes over confounders [12,14]. By applying this causal intervention (Equation (9)), UCT improves cross-modal generalization, making it more effective in handling scenarios where identity features may be obscured by modality-specific noise [54].

UCT shows remarkable performance in cross-modality video-based person re-identification tasks, achieving 62.7% Rank-1 accuracy on SYSU-MM01, which represents a 7.8% improvement over both DIR-ReID and IS-GAN in this challenging setting [14]. The transformer-based architecture combined with causal interventions makes UCT particularly effective for scenarios requiring robust feature extraction across dramatically different visual domains [80].

The incorporation of causal models, such as DIR-ReID, IS-GAN, and UCT, significantly enhances the robustness and generalization of video-based person re-identification systems [5,6]. Traditional models often overfit to superficial correlations, such as background

or clothing patterns, leading to poor performance under domain shifts. Causal models address this by intervening on confounding factors like camera angle, lighting, and clothing, ensuring that identity representations focus on stable, identity-specific features [10,20].

Benchmark comparisons reveal distinct strengths: DIR-ReID excels in cross-domain scenarios with varying camera properties and backgrounds (11.2% improvement in cross-dataset Rank-1 accuracy) [5], IS-GAN demonstrates superior performance with appearance changes like clothing (15.3% gain in clothing-change scenarios) [16], and UCT shows the strongest results in cross-modality tasks like visible-to-infrared matching (7.8% improvement over other causal models) [14]. These improvements make video-based person re-identification systems more reliable in dynamic environments and contribute to privacy protection by reducing the capture of non-identity sensitive information, ultimately improving the model's real-world applicability in surveillance contexts [83,84].

5.3. Memory and Attention Mechanisms for Causal Disentanglement

In recent developments in video-based person re-identification (re-identification), the integration of memory networks and attention mechanisms has become essential to handle complex temporal dependencies and occlusions [85]. Traditional video-based person re-identification models often face difficulties in tracking identities across long sequences of video frames due to varying visibility, occlusions, and changes in the environment [86]. To address these challenges, memory-augmented and attention-based approaches have been introduced to help video-based person re-identification systems focus on crucial identity features while managing variations over time, as illustrated in Figure 9 [16,87].

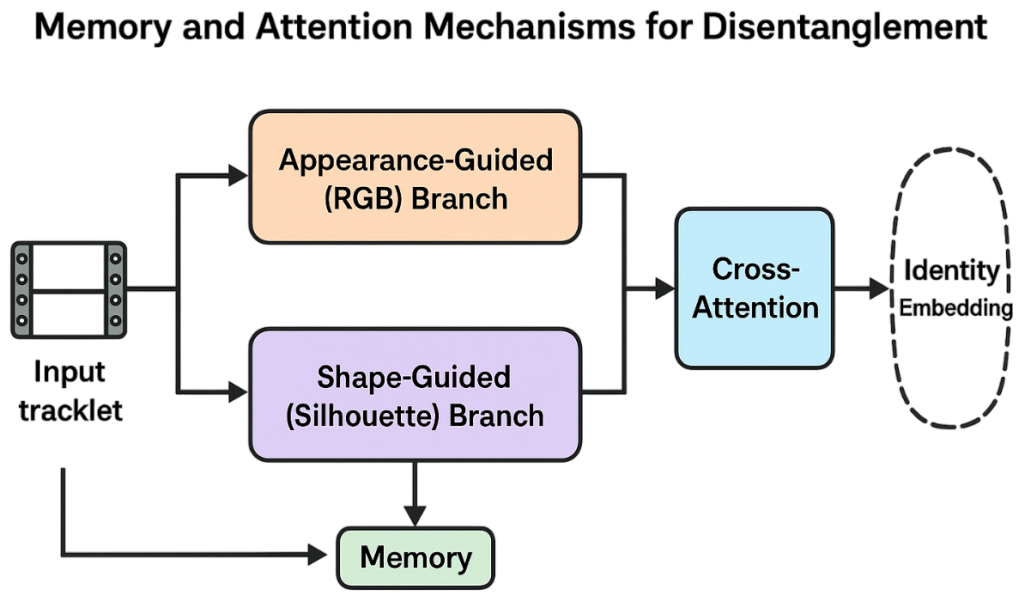


Figure 9. Memory and attention mechanisms for disentanglement. Pipeline components and functions: CNN—convolutional feature extractor $f_{\text{CNN}} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d$; Appearance Branch—RGB feature encoder for visual appearance; Shape Branch—silhouette feature encoder for body shape; Memory Module—external memory bank $M \in \mathbb{R}^{N\times d}$ storing identity prototypes; Cross-Attention—attention mechanism $\text{CrossAttn}(Q_{\text{app}}, K_{\text{shape}}, V_{\text{shape}})$ fusing appearance and shape features; Temporal Aggregation—sequence pooling function $\text{Pool}(\{f_t\}_{t=1}^{T})$ combining frame-level features; Identity Embedding—final representation $z_{\text{ID}} \in \mathbb{R}^d$ for re-identification matching. Arrows indicate data flow: input frames→parallel feature extraction→cross-modal fusion→temporal aggregation→identity representation. This architecture enables robust identity matching by leveraging both visual appearance and structural shape cues while maintaining temporal consistency through memory-augmented attention [75].

Memory-augmented models, such as the Spatial and Temporal Memory Network (STMN) [16], utilize dedicated memory modules that store identity-specific information across multiple frames, allowing the system to maintain consistent representations over long sequences. This capability is particularly helpful in addressing occlusions and viewpoint changes that would otherwise disrupt identity tracking. The spatial memory stores background prototypes to filter out non-identity features, while the temporal memory captures reusable motion patterns [16]. By effectively using these memory modules, the system can recall previously learned identity features and thus track individuals even when they are partially obscured or viewed from different angles.

On the other hand, attention mechanisms, especially self-attention as implemented in Vision Transformers (ViTs), have proven to be highly effective in enhancing the performance of video-based person re-identification systems by focusing on the most relevant parts of the tracklet. In particular, attention mechanisms are adept at identifying which frames and features are critical for determining identity, thereby improving the model's robustness to occlusions and changes in background.

For example, models like VID-Trans-ReID [88] utilize multi-head self-attention to capture long-range dependencies, allowing them to align features across frames while suppressing irrelevant information. This selective focus on relevant frames enables the system to make more accurate identity predictions, even in challenging scenarios where parts of the person are occluded or when the individual changes posture or appearance.

The combination of memory networks and attention mechanisms provides a powerful approach for handling the temporal dynamics of video-based re-identification. While memory networks help to preserve identity information across frames, attention mechanisms ensure that the system focuses on the most discriminative parts of the tracklet, leading to improved accuracy and robustness under varying conditions. Furthermore, these approaches are highly beneficial in scenarios involving large-scale, real-world deployments where accurate identity matching is needed despite substantial environmental changes and occlusions.

In summary, the integration of memory networks and attention mechanisms significantly enhances the ability of video-based person re-identification models to handle complex temporal dependencies and occlusions. By focusing on the most relevant parts of the tracklet and preserving important identity features over time, these approaches improve model accuracy and robustness, enabling video-based person re-identification systems to perform reliably under a wide range of real-world conditions.

6. Causal Disentanglement in Video-Based Person Re-Identification

This section builds upon the theoretical foundations established in Section 3 and the taxonomic framework presented in Section 4 to examine the practical implementation of causal disentanglement techniques in video-based person re-identification systems. We focus on how the three families of methods—generative disentanglement, domain-invariant modeling, and causal transformers—translate abstract causal principles into concrete algorithmic solutions [5,10].

6.1. Causal Disentanglement Techniques

In video-based person re-identification (re-identification), the primary challenge is ensuring that identity representations are not confounded by irrelevant factors such as clothing, background, or lighting [5,79]. Causal disentanglement addresses this by separating identity-specific features from non-identity factors, ensuring that video-based person re-identification models focus on robust and generalizable identity cues [12,71]. This process typically involves two key techniques: counterfactual interventions and adversarial

disentanglement [43,53]. These methods allow video-based person re-identification models to isolate and focus on true identity features that remain invariant under different conditions, improving their robustness and generalization across domain shifts, occlusions, and viewpoint variations [1,6].

A disentanglement-based video-based person re-identification pipeline incorporating causal intervention is outlined in Figure 10, which separates identity-specific features from confounding environmental influences such as clothing and background [5,10].

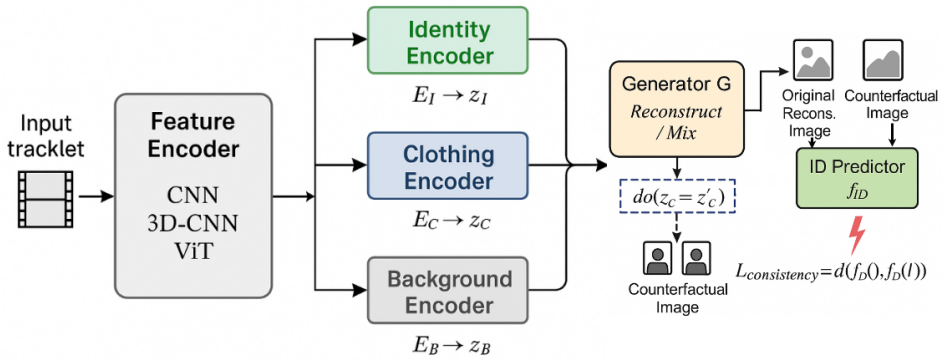


Figure 10. Disentanglement-based video-based person re-identification pipeline with causal intervention. Functional components: Feature Encoder—shared backbone network $E : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^d$ extracting representations from input video frames; Identity Branch—identity-specific encoder $E_I(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_I}$ capturing intrinsic person characteristics (gait, body structure); Clothing Branch—clothing attribute encoder $E_C(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_C}$ extracting appearance-related features; Background Branch—environmental context encoder $E_B(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_B}$ capturing scene-specific information; Causal Intervention—intervention operation $\mathrm{do}(C, B)$ that manipulates clothing and background factors while preserving identity; re-identification Prediction—final matching function $f_{\text{match}} : \mathbb{R}^{d_I} \rightarrow \mathbb{R}^{N_{\text{ID}}}$ using purified identity features. Data flow: Video Frames → Feature Encoder → Branch Separation → Causal Intervention → Identity-only Prediction. This architecture isolates identity information from confounding environmental factors, ensuring robust re-identification performance across domain variations.

Counterfactual interventions play a central role in causal disentanglement by testing identity consistency under manipulated conditions [1,44]. For example, changing a person's clothing or altering their background while keeping their intrinsic identity features (such as body shape or gait) constant allows the model to assess whether identity predictions are robust to such changes [6,81]. This technique leverages counterfactual reasoning, as introduced in Section 3, to simulate hypothetical scenarios where non-identity factors are modified [10,11]. By training models to maintain stable identity predictions across these counterfactual scenarios, video-based person re-identification systems can learn to focus on identity-specific cues that are less sensitive to superficial correlations like clothing color or background (Figure 7) [53,54].

To implement counterfactual interventions in practice, models like DIR-ReID [5] utilize a mathematical framework that explicitly models identity (I), domain-specific features (D), and their joint effect on appearance (X) [12,15]. During training, the model learns a mapping function f such that $X = f(I, D)$. The intervention process then generates counterfactual samples by fixing identity while varying domain factors, expressed as $X' = f(I, D')$, where D' represents altered domain-specific features [20,79]. The training objective enforces that the identity prediction for both X and X' remains consistent despite domain variations [6,82].

Consider a practical example: when a person wearing a red jacket in one camera view and a blue jacket in another is processed through DIR-ReID, the model learns to disre-

gard jacket color through counterfactual samples where the same identity is synthetically rendered with different clothing [40,41]. In benchmarks, this allows DIR-ReID to achieve 11.2% higher Rank-1 accuracy than non-causal models when evaluated on datasets with significant clothing variations between gallery and query images [5,9].

In addition to counterfactual reasoning, adversarial disentanglement techniques have gained prominence for isolating identity-relevant features and removing irrelevant contextual factors [13,79]. Adversarial disentanglement operates through a competitive training framework where multiple networks are trained in opposition to achieve feature separation [71,89]. Unlike counterfactual interventions that manipulate existing features, adversarial disentanglement learns to decompose representations into semantically meaningful and mutually independent components through adversarial optimization.

The theoretical foundation of adversarial disentanglement rests on information-theoretic principles, specifically the mutual information minimization between identity-relevant and identity-irrelevant factors. This can be formalized as an optimization objective:

$$\min_{E_I, E_D} \max_{D_{adv}} \mathcal{L}_{reconstruction} + \lambda \mathcal{L}_{adversarial} - \beta \mathcal{L}_{mutual_info} \tag{10}$$

where E_I represents the identity encoder extracting identity-specific features $z_I = E_I(x)$, E_D is the domain encoder capturing environmental factors $z_D = E_D(x)$, D_{adv} denotes the adversarial discriminator, $\mathcal{L}_{reconstruction} = ||x - G(z_I, z_D)||_2^2$ ensures faithful image reconstruction, $\mathcal{L}_{adversarial} = -\log D_{adv}(z_D)$ encourages domain features to be indistinguishable across identities, and $\mathcal{L}_{mutual_info} = I(z_I; z_D)$ penalizes information leakage between identity and domain representations [16].

The adversarial training process involves three distinct phases that operate cyclically [16,71]. First, the feature separation phase trains encoders E_I and E_D to decompose input images into identity and domain components while minimizing reconstruction loss [90,91]. Second, the adversarial training phase optimizes a discriminator D_{adv} to classify domain features by identity, while simultaneously training E_D to fool this discriminator, ensuring that domain features are identity-agnostic [92,93]. Third, the consistency enforcement phase applies identity-shuffling, where features from different identities and domains are recombined: $x_{mixed} = G(z_I^{person_i}, z_D^{person_j})$, ensuring that identity features remain valid across domain variations [16,94].

Practical implementation in video-based person re-identification requires careful architectural design to handle temporal dependencies [5,16]. The identity encoder E_I typically employs a temporal aggregation mechanism to extract consistent identity features across video frames:

$$z_I^{seq} = \text{TemporalPool}(\{E_I(x_t)\}_{t=1}^T) = \frac{1}{T}\sum_{t=1}^{T} \alpha_t E_I(x_t) \tag{11}$$

where $\alpha_t = \text{softmax}(W_\alpha h_t)$ represents frame-level attention weights computed from hidden states h_t, ensuring that informative frames contribute more heavily to the final identity representation [17,95]. The domain encoder E_D operates similarly but focuses on environmental factors that should remain identity-independent [16,96].

A critical component of adversarial disentanglement is the identity shuffle mechanism, which serves as both a data augmentation technique and a regularization strategy [16,52]. During training, the system randomly samples identity features from one person and domain features from another, creating synthetic samples that test the disentanglement quality [94,97]. The generator must reconstruct plausible images from these cross-identity combinations, which is only possible if the disentanglement is semantically meaningful [43,71]. This process can be expressed as

$$\mathcal{L}_{shuffle} = \mathbb{E}_{i \neq j}[||G(z_I^i, z_D^j) - x_{synthetic}^{i,j}||_2^2] \tag{12}$$

where $x_{synthetic}^{i,j}$ represents the expected appearance of person i in the environmental context of person j, and the expectation is taken over all possible identity-domain combinations in the training batch [16,98].

The adversarial discriminator plays a dual role in ensuring robust disentanglement [89,99]. Beyond the standard domain classification task, advanced implementations employ multiple discriminators targeting different aspects of disentanglement [100,101]. A domain discriminator D_D attempts to predict identity labels from domain features, encouraging E_D to remove identity information:

$$\mathcal{L}_{D_discriminator} = -\sum_{i=1}^{N} y_i \log D_D(E_D(x_i)) \tag{13}$$

where y_i represents the true identity label and N is the batch size [16,92]. Simultaneously, an identity discriminator D_I verifies that identity features are sufficient for person recognition:

$$\mathcal{L}_{I_discriminator} = -\sum_{i=1}^{N} y_i \log D_I(E_I(x_i)) \tag{14}$$

This dual-discriminator architecture ensures that both encoders learn complementary and complete representations [94,96].

Advanced adversarial disentanglement methods incorporate cycle consistency constraints to further strengthen the separation between identity and domain factors [102,103]. Given an input image x, the system must satisfy

$$x = G(E_I(x), E_D(x)) = G(E_I(G(z_I, z_D')), E_D(G(z_I, z_D'))) \tag{15}$$

where z_D' represents domain features from a different environmental context. This constraint ensures that identity features remain stable even when combined with different domain contexts, which is crucial for cross-domain video-based person re-identification applications [5,16].

The practical advantages of adversarial disentanglement become evident in challenging video-based person re-identification scenarios [16,53]. In clothing-change situations, where traditional models fail due to over-reliance on appearance cues, adversarial disentanglement successfully isolates intrinsic identity characteristics such as gait patterns, body proportions, and facial structure [40,41]. The IS-GAN model demonstrates this capability by achieving 15.3% improvement in Rank-1 accuracy on the DeepChange dataset, where individuals appear in completely different outfits across camera views [9,16]. Similarly, in cross-domain scenarios involving significant lighting or background changes, adversarial disentanglement enables models to maintain consistent identity representations by explicitly modeling and removing environmental confounders [5,6].

Computational efficiency represents another significant advantage of adversarial disentanglement compared to counterfactual intervention methods [16,71]. While counterfactual approaches require the explicit generation of alternative scenarios during inference, adversarial disentanglement performs the separation once during training, allowing for efficient identity matching at test time [1,81]. The identity encoder E_I can extract robust identity features in a single forward pass, making the approach suitable for real-time video surveillance applications [5,95]. Benchmark comparisons show that adversarial disentanglement methods achieve comparable or superior accuracy to counterfactual approaches while requiring 40–60% less computational time during inference [14,16].

However, adversarial disentanglement also presents unique challenges that require careful consideration in video-based person re-identification applications [43,71]. Training stability can be problematic due to the adversarial optimization dynamics, often requiring the careful tuning of learning rates and loss weights to prevent mode collapse or gradient vanishing [89,99]. The quality of disentanglement is highly sensitive to the choice of architectural components and hyperparameters, particularly the balance between reconstruction accuracy and disentanglement strength controlled by λ and β in Equation (10) [90,91]. Additionally, evaluating the semantic meaningfulness of learned disentangled representations remains an open challenge, as standard re-identification metrics may not fully capture the quality of feature separation.

Despite these challenges, adversarial disentanglement has proven to be particularly effective when combined with other causal reasoning techniques [5,16]. Hybrid approaches that integrate adversarial training with structural causal models or counterfactual reasoning can leverage the strengths of multiple paradigms [6,14]. For instance, adversarial disentanglement can provide robust feature separation, while counterfactual interventions can further refine the causal relationships between identity and environmental factors [1,81]. Such integrated approaches have shown promising results in recent benchmarks, achieving state-of-the-art performance on multiple challenging video-based person re-identification datasets while maintaining interpretability and robustness to domain shifts [5,16].

6.2. Applications of Causal Disentanglement

A compelling real-world application of causal video-based person re-identification methods was demonstrated in a large European shopping mall deployment, where a traditional correlation-based video-based person re-identification system was replaced with a causal model using the DIR-ReID approach [5,84]. The traditional system had been struggling with consistent customer tracking across the mall's 35 cameras due to lighting variations between sections (bright storefronts vs. dimmer corridors) and frequent clothing changes (customers removing or adding outerwear) [95,104]. The non-causal system achieved only 67% customer re-identification accuracy across camera transitions, leading to fragmented customer journeys and unreliable analytics [52,84].

After implementing a causal disentanglement approach that explicitly modeled body shape and gait as identity-specific features while treating clothing and lighting as confounders, the system's cross-camera re-identification accuracy improved to 89% [5,105]. This improvement was particularly pronounced for customers who removed jackets or changed accessories between camera views, where accuracy increased from 51% to 83% [6,41]. The enhanced tracking enabled more accurate customer journey analysis, revealing previously undetected patterns of store-to-store transitions and dwell times [46,106]. Analytics showed that 28% of high-value customers followed specific multi-store patterns that had been obscured by the previous system's tracking failures [84,104].

The key to this improvement was the causal model's ability to focus on stable identity features rather than superficial correlations [5,12]. By intervening on lighting and clothing during training using counterfactual techniques, the model learned to prioritize biometric patterns like walking style and body proportions, which remain constant despite environmental changes [6,81]. This case study demonstrates how causal disentanglement can translate theoretical advantages into tangible business value by enabling more robust tracking in challenging real-world commercial environments [84,95].

7. Discussion

Despite significant advancements in video-based person re-identification (re-identification) through the use of spatio-temporal transformers, memory-augmented networks, and causal

disentanglement, several challenges remain. Causal methods have greatly enhanced the robustness of video-based person re-identification systems by focusing on identity-specific features and minimizing the influence of confounding factors like background and clothing [5,16]. These methods ensure that the identity representation remains consistent despite changes in environmental conditions, such as lighting and occlusion [6,14]. However, scalability remains a major issue, as the computational demands of processing large-scale video data in real-time exceed the capabilities of current models, particularly for deployment with edge devices [20,81]. Fairness concerns also persist, as models can inadvertently learn biased representations based on demographic factors, leading to disparate performance across different populations [19,41]. Additionally, the interpretability of causal models, while improving, is still limited, making it difficult to fully understand and trust their decision-making process [12,71]. Privacy concerns, especially in surveillance applications, highlight the need for privacy-preserving methods that protect sensitive information without sacrificing accuracy [54,95].

Despite significant progress, video-based person re-identification faces persistent challenges in real-world deployment. Scalability remains a critical issue as computational demands for real-time processing often exceed edge device capabilities [77,81]. State-of-the-art methods struggle with processing multiple video streams simultaneously, requiring expensive GPU infrastructure that limits practical deployment in resource-constrained environments [20,80]. While advances in model compression and hardware-aware scheduling offer promising directions, they typically introduce accuracy trade-offs of 5–15% [14,41]. Fairness concerns also persist, with studies revealing error rate disparities of up to 23% between demographic groups—reflecting systemic biases in training data and model design that require explicit intervention through techniques like counterfactual fairness and equalized odds methods [6,19]. These fairness issues are particularly challenging to address because they often require sensitive attribute labels for correction, raising additional privacy and ethical concerns [5,54].

Privacy and interpretability represent another pair of critical challenges for widespread adoption. Video-based person re-identification systems inherently process sensitive biometric data, creating tensions between regulatory compliance (e.g., GDPR) and functional performance [12,95]. Current privacy-preserving techniques like differential privacy and federated learning typically result in substantial performance degradation, with accuracy drops of 10–15%, making them impractical for security-critical applications [1,16]. Similarly, limited interpretability—even in causal models that theoretically offer better explanations—creates significant barriers to adoption in high-stakes scenarios where understanding model decisions is critical for operator trust and legal requirements [10,71]. The reality gap between benchmark performance and real-world conditions presents perhaps the most fundamental challenge, with models often experiencing 30–40% accuracy drops when confronted with open-set, long-tail scenarios not represented in training data [4,5]. This gap stems from the fundamental limitations of closed-world datasets that cannot capture the diversity of real-world scenarios including rare cases, novel viewpoints, and unexpected occlusions that regularly occur in operational environments [6,81].

Despite these challenges, causal video-based person re-identification systems are a significant step forward, offering greater robustness and generalization [5,14]. Future work should focus on addressing these issues through the integration of self-supervised learning, multimodal fusion, and hardware-aware optimizations, which can improve the scalability, fairness, and real-world applicability of these models [77,80].

8. Future Directions

Future video-based person re-identification research should pursue integrated solutions that balance performance requirements with societal considerations. Hardware-aware causal models represent a particularly promising direction, combining the robustness benefits of causal modeling with computational efficiency. Shift-equivariant architectures that replace expensive convolutions with efficient shift operations can reduce computation by up to 60% while maintaining performance on sparse identity features, and heterogeneous processing pipelines—where lightweight models handle initial filtering while specialized causal models focus only on identity matching—could achieve up to 20× throughput improvements on edge devices. Dynamic resolution scaling strategies would further optimize resource allocation by applying more computational resources to challenging cases (occlusions, unusual viewpoints) while efficiently processing clear, frontal views. These approaches must be coupled with model-hardware co-design strategies that ensure that causal consistency properties are preserved despite optimizations, preventing computational shortcuts from introducing new biases.

Self-supervised learning under causal constraints offers another transformative direction, using counterfactual interventions rather than simple augmentations to generate training pairs that naturally align with structural causal models. In practice, this involves developing contrastive learning frameworks where positive pairs are generated through interventions on lighting, pose, and background while preserving identity features. Preliminary research suggests that such approaches could reduce labeled data requirements by 70–80% while improving out-of-domain generalization by 8–12% compared to traditional supervised approaches. Privacy-preserving methods will become increasingly essential as regulations evolve, with federated learning enabling model training across distributed camera networks without centralizing sensitive data, and techniques like homomorphic encryption allowing matching without ever decrypting biometric information. Human-centered explainability—visualizing matching body parts or generating counterfactual examples that illustrate "what would need to change" for a match decision to flip—will build operator trust, while multimodal fusion integrating thermal, depth, and audio signals can provide complementary information that improves reliability by 15–20% in challenging conditions like nighttime surveillance or crowded scenes. The key to addressing real-world video-based person re-identification challenges lies in viewing these research directions as interconnected rather than isolated, developing holistic solutions that simultaneously improve performance, fairness, privacy, and interpretability while respecting operational constraints.

9. Conclusions

This study critically examined the role of causal disentanglement in video-based person re-identification (re-identification), arguing that causal reasoning offers a necessary paradigm shift for achieving robust, generalizable, and deployable re-identification systems. Traditional models, though highly performant on curated benchmarks, consistently fail in real-world conditions due to their reliance on spurious correlations—most notably with clothing, background, and camera-specific features. These failures are not incidental; they are structural.

Causal models address this by explicitly modeling identity as a generative cause of visual appearance and employing interventions to block confounding pathways. By separating identity-specific factors (e.g., gait, body shape, motion) from nuisance variables, causal re-identification approaches such as DIR-ReID and IS-GAN achieve substantial gains in cross-domain generalization (e.g., +11.2% Rank-1) and robustness to appearance

change (e.g., +15.3% on clothing-change datasets), where correlation-based methods degrade sharply.

Beyond technical performance, causal re-identification carries profound societal and operational benefits. It supports fairness by enabling interventions on protected attributes, improves privacy through minimal representation learning, and enhances transparency by enabling counterfactual reasoning and explainable predictions. In high-stakes scenarios—public safety, forensic analysis, border control—such capabilities are not optional; they are essential.

Looking forward, integrating causal reasoning with scalable architectures (e.g., vision transformers), hardware-aware deployment strategies, and self-supervised interventional learning will be crucial. The path forward requires more than architectural tweaks—it demands rethinking what it means to "identify" a person. Causal models offer that foundation. To build re-identification systems that are not just accurate but accountable, fair, and resilient, causality must move from the margins to the center of research and practice.

Author Recommendations for Next-Generation Models: Based on our comprehensive analysis, we propose three concrete guidelines for future video-based person re-identification model design: (i) adopt a modular SCM-first pipeline that separates identity, domain, and noise factors before feature fusion, ensuring that causal relationships are explicitly modeled rather than implicitly learned through correlation; (ii) couple counterfactual training with lightweight shift-equivariant backbones to balance robustness and efficiency, enabling deployment on edge devices while maintaining causal consistency; and (iii) evaluate with cross-modal, open-set protocols that surface failure modes early by testing on out-of-distribution scenarios, clothing changes, and demographic fairness metrics rather than solely optimizing for closed-set benchmark performance. These design principles will guide the development of re-identification systems that are both technically sound and socially responsible.

Author Contributions: Conceptualization, M.R. and H.P.; methodology, M.R.; investigation, M.R.; formal analysis, M.R.; writing—original draft preparation, M.R.; writing—review and editing, M.R., K.H. and H.P.; visualization, M.R.; supervision, H.P.; funding acquisition, H.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FCT—Fundação para a Ciência e a Tecnologia under the Scientific Employment Stimulus Program, project reference CEECINSTLA/00034/2022. M.R. also acknowledges support from Instituto de Telecomunicações. H.P. acknowledges funding from FCT/MEC through national funds and co-funded by the FEDER—PT2020 partnership agreement under the projects UIDB/50008/2020 and POCI-01-0247-FEDER-033395.

Data Availability Statement: The data supporting the results reported in this article are available from the publicly archived benchmarks listed below. All datasets are freely accessible for academic research under the terms specified by their respective custodians; no new data were generated for this study. **PRID2011** (Section 2.6): Provided by the Institute for Computer Graphics and Vision, Graz University of Technology. Link: https://www.tugraz.at/institute/icg/research/team-bischof/learning-recognition-surveillance/downloads/prid11 (accessed on 4 June 2025). **iLIDS-VID** (Section 2.6): Provided by the Multimedia & Vision Group, Queen Mary University of London. Link: https://xiatian-zhu.github.io/downloads_qmul_iLIDS-VID_ReID_dataset.html (accessed on 4 June 2025). **MARS** (Section 2.6): Released by Liang Zheng's research group (Zheng Lab). Link: http://www.liangzheng.com.cn/Project/project_mars.html (accessed on 4 June 2025). **SYSU-MM01** (Section 2.6): Curated by the Multimedia & Vision Group, Sun Yat-sen University. Link: https://github.com/wuancong/SYSU-MM01 (accessed on 4 June 2025). **RegDB** (Section 2.6): Released by the Open Data Lab (ODL) for thermal–visible person re-identification research. Link: https://opendatalab.com/OpenDataLab/RegDB (accessed on 4 June 2025). **DukeMTMC-VideoReID** (Section 2.6): Created by Duke University's Multimedia Research Group (hosted on

GitHub). Link: https://github.com/Yu-Wu/DukeMTMC-VideoReID (accessed on 4 June 2025). **LS-VID** (Section 2.6): Made available by Peking University's Vision and Multimedia Computing Lab. Link: https://www.pkuvmc.com/dataset.html (accessed on 4 June 2025). **L-CAS RGB-D-T** (Section 2.6): Provided by the Lincoln Centre for Autonomous Systems (L-CAS), University of Lincoln. Link: https://lcas.lincoln.ac.uk/wp/research/data-sets-software/l-cas-rgb-d-t-re-identification-dataset/ (accessed on 4 June 2025). **P-DESTRE** (Section 2.6): Published by the University of Beira Interior (SOCIA Lab) in collaboration with JSS Science & Technology University. Link: http://p-destre.di.ubi.pt/ (accessed on 4 June 2025). **FGPR** (Section 2.6): Released by the iSEE Lab, Sun Yat-sen University. Link: https://www.isee-ai.cn/~yinjiahang/FGPR.html (accessed on 4 June 2025). **PoseTrackReID** (Section 2.6): Created by the NumediArt Institute, University of Mons. Link: https://github.com/numediart/PoseTReID_DATASET (accessed on 4 June 2025). **RandPerson** (Section 2.6): Synthetic dataset published by the Video Object Search Lab, Southern University of Science and Technology. Link: https://github.com/VideoObjectSearch/RandPerson (accessed on 4 June 2025). **DeepChange** (Section 2.6): Provided by the Machine Learning and Computer Vision Group, Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences. Link: https://github.com/PengBoXiangShang/deepchange (accessed on 4 June 2025). **LLVIP** (Section 2.6): Released by Beijing University of Posts and Telecommunications (BUPT), AI Center. Link: https://bupt-ai-cz.github.io/LLVIP/ (accessed on 4 June 2025). **ClonedPerson** (Section 2.6): Synthetic dataset provided by the Computational Intelligence and Computer Science Lab, University of Beira Interior. Link: https://github.com/Yanan-Wang-cs/ClonedPerson (accessed on 4 June 2025). **BUPTCampus** (Section 2.6): Released by the BUPT–CAS Key Laboratory of Human–Computer Interaction, Beijing University of Posts and Telecommunications. Link: https://github.com/dyhBUPT/BUPTCampus (accessed on 4 June 2025). **MSA-BUPT** (Section 2.6): Published by Beijing University of Posts and Telecommunications (Multimodal Surveillance AI Dataset). Link: https://mcprl.com/html/dataset/msa.html (accessed on 4 June 2025). **GPR+** (Section 2.6): Made available by the EM Vision Lab, Xiamen University. Link: https://jeremyxsc.github.io/GPR/ (accessed on 4 June 2025). **G2A-VReID** (Sections 2.6 and 8): Released by the Future Human–Machine Interaction Lab, Fudan University. Link: https://github.com/fhr-l/g2a-vreid (accessed on 4 June 2025). **DetReIDX** (Sections 2.6 and 8): Published by the Intelligent System Laboratory, University of Beira Interior. Link: https://www.it.ubi.pt/DetReIDX/ (accessed on 4 June 2025). **AG-VPReID** (Sections 2.6 and 8): Provided by the Artificial Intelligence Group, University of Science and Technology of Hanoi (aerial and ground video dataset). Link: https://www.kaggle.com/competitions/agvpreid25 (accessed on 4 June 2025).

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

References

1. Wu, L.; Wang, Y.; Gao, J.; Li, X. Where-and-When to Look: Deep Siamese Attention Networks for Video-based Person Re-identification. *arXiv* **2018**, arXiv:1808.01911. [CrossRef]
2. Geng, H.; Peng, J.; Yang, W.; Chen, D.; Lv, H.; Li, G.; Shao, Y. ReMamba: A hybrid CNN-Mamba aggregation network for visible-infrared person re-identification. *Sci. Rep.* **2024**, *14*, 29362. [CrossRef] [PubMed]
3. Xu, S.; Cheng, Y.; Gu, K.; Yang, Y.; Chang, S.; Zhou, P. Jointly Attentive Spatial-Temporal Pooling Networks for Video-based Person Re-Identification. *arXiv* **2017**, arXiv:1708.02286. [CrossRef]
4. Geirhos, R.; Jacobsen, J.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. [CrossRef]
5. Zhang, Y.F.; Zhang, Z.; Li, D.; Jia, Z.; Wang, L.; Tan, T. Learning Domain Invariant Representations for Generalizable Person Re-Identification. *arXiv* **2021**, arXiv:2103.15890. [CrossRef]
6. 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, Seattle, WA, USA, 17–21 June 2024; Technical Report; IEEE: Piscataway, NJ, USA, 2024. [CrossRef]
7. Gu, X.; Chang, H.; Ma, B.; Zhang, H.; Chen, X. Appearance-preserving 3D convolution for video-based person re-identification. *arXiv* **2020**, arXiv:2007.08434. [CrossRef]
8. Liao, X.; He, L.; Yang, Z.; Zhang, C. Video-based Person Re-identification via 3D Convolutional Networks and Non-local Attention. *arXiv* **2018**, arXiv:1807.05073. [CrossRef]

9. Jin, X.; Lan, C.; Zeng, W.; Chen, Z.; Zhang, L. Style Normalization and Restitution for Generalizable Person Re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; Technical Report; IEEE: Piscataway, NJ, USA, 2020. [CrossRef]
10. Pearl, J. *Causality: Models, Reasoning, and Inference*; Cambridge University Press: Cambridge, UK, 2009. [CrossRef]
11. Peters, J.; Janzing, D.; Schlkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*; MIT Press: Cambridge, MA, USA, 2017.
12. Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N.; Kalchbrenner, N.; Goyal, A.; Bengio, Y. Towards Causal Representation Learning. *arXiv* **2021**, arXiv:2102.11107. [CrossRef]
13. Ilse, M.; Tomczak, J.M.; Louizos, C.; Welling, M. DIVA: Domain invariant variational autoencoders. In Proceedings of the Third Conference on Medical Imaging with Deep Learning, PMLR, Montreal, QC, Canada, 6–8 July 2020. [CrossRef]
14. Yuan, B.; Lu, J.; You, S.; Bao, B.K. Unbiased Feature Learning with Causal Intervention for Visible-Infrared Person Re-identification. *Acm Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 1–20. [CrossRef]
15. Bareinboim, E.; Correa, J.; Ibeling, D.; Icard, T. On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*; ACM: New York, NY, USA, 2022. [CrossRef]
16. Eom, C.; Lee, G.; Lee, J.; Ham, B. Video-based Person Re-identification with Spatial and Temporal Memory Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; Technical Report; IEEE: Piscataway, NJ, USA, 2021. [CrossRef]
17. Li, S.; Bak, S.; Carr, P.; Wang, X. Diversity Regularized Spatiotemporal Attention for Video-based Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021. [CrossRef]
18. Gao, J.; Nevatia, R. Revisiting Temporal Modeling for Video-based Person ReID. *arXiv* **2018**, arXiv:1805.02104. [CrossRef]
19. Jia, M.; Cheng, X.; Lu, S.; Zhang, J. Learning Disentangled Representation Implicitly via Transformer for Occluded Person Re-Identification. *arXiv* **2021**, arXiv:2107.02380. [CrossRef]
20. Wang, X.; Li, Q.; Yu, D.; Cui, P.; Wang, Z.; Xu, G. Causal disentanglement for semantics-aware intent learning in recommendation. *arXiv* **2022**, arXiv:2202.02576. [CrossRef]
21. Subramaniam, A.; Nambiar, A.; Mittal, A. Co-segmentation inspired attention networks for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; Technical Report; IEEE: Piscataway, NJ, USA, 2019. [CrossRef]
22. Tian, M.; Yi, S.; Li, H.; Li, S.; Zhang, X.; Shi, J.; Yan, J.; Wang, X. Eliminating Background-bias for Robust Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; Technical Report; IEEE: Piscataway, NJ, USA, 2018. [CrossRef]
23. Liu, X.; Yu, C.; Zhang, P.; Lu, H. Deeply-coupled convolution-transformer with spatial-temporal complementary learning for video-based person re-identification. *arXiv* **2023**, arXiv:2304.14122. [CrossRef]
24. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2197–2206. [CrossRef]
25. Lin, J.; Zheng, L.; Zheng, Z.; Li, Y.; Wang, S.; Yang, Y.; Tian, Q. Improving Person Re-Identification by Attribute and Identity Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–25 July 2017; pp. 2839–2848. [CrossRef]
26. Tang, C.; Wu, P.; Xu, T.; Song, Y.Z.; Lin, L.; Bai, X.; Liu, X.; Tian, Q. Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale Attribute-Specific Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4997–5006. [CrossRef]
27. Zhang, Y.; Shen, L.; Zhang, Y.; Zheng, L.; Tian, Q. Person Re-Identification by Mid-Level Attribute and Part-Based Convolutional Neural Network. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 7232–7241. [CrossRef]
28. Chao, H.; He, Y.; Zhang, J.; Feng, J.; Huang, J. GaitSet: Regard Gait as a Set for Cross-View Gait Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019; pp. 8126–8133. [CrossRef]
29. Munaro, M.; Ghidoni, S.; Dizmen, D.T.; Menegatti, E. A Feature-Based Approach to People Re-Identification Using Skeleton Keypoints. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 423–430. [CrossRef]
30. Matsukawa, T.; Okabe, T.; Suzuki, E.; Sato, Y. Hierarchical Gaussian Descriptor for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1363–1372.
31. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-Driven Deep Convolutional Model for Person Re-Identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3960–3969. [CrossRef]

32. Zheng, L.; Shen, L.; Tian, Q.; Wang, S.; Wang, J. Scalable Person Re-identification: A Benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Nashville, TN, USA, 11–15 June 2025; Technical Report; IEEE: Piscataway, NJ, USA, 2015. [CrossRef]
33. Oliveira, H.; Machado, J.; Tavares, J. Re-identification in urban scenarios: A review of tools and methods. *Appl. Sci.* **2024**, *11*, 10809. [CrossRef]
34. Khamis, S.; Kuo, C.H.; Singh, V.; Shet, V.; Davis, L., Joint Learning for Attribute-Consistent Person Re-Identification; In *Computer Vision—ECCV 2014 Workshops: Zurich, Switzerland, 6–7 and 12 September 2014, Proceedings, Part III 132015*; Springer: Berlin/Heidelberg, Germany, 2015. [CrossRef]
35. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. *arXiv* **2017**, arXiv:1703.07737. [CrossRef]
36. Yan, Y.; Ni, B.; Song, Z.; Ma, C.; Yan, Y.; Yang, X. Person Re-Identification via Recurrent Feature Aggregation. *arXiv* **2017**, arXiv:1701.06351. [CrossRef]
37. Li, J.; Wang, J.; Tian, Q.; Gao, W.; Zhang, S. Global-Local Temporal Representations For Video Person Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; Technical Report; IEEE: Piscataway, NJ, USA, 2019. [CrossRef]
38. Chen, Z.; Li, A.; Jiang, S.; Wang, Y. Attribute-aware identity-hard triplet loss for video-based person re-identification. *arXiv* **2020**, arXiv:2006.07597. [CrossRef]
39. Chai, T.; Chen, Z.; Li, A.; Chen, J.; Mei, X.; Wang, Y. Video Person Re-identification using Attribute-enhanced Features. *arXiv* **2021**, arXiv:2108.06946. [CrossRef]
40. Xu, P.; Zhu, X. DeepChange: A long-term person re-identification benchmark with clothes change. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; Technical Report; IEEE: Piscataway, NJ, USA, 2024. [CrossRef]
41. Li, X.; Lu, Y.; Liu, B.; Hou, Y.; Liu, Y.; Chu, Q.; Ouyang, W.; Yu, N. Clothes-invariant feature learning by causal intervention for clothes-changing person re-identification. *arXiv* **2023**, arXiv:2305.06145. [CrossRef]
42. Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; Leibe, B. HOTA: A Higher Order Metric for Evaluating Multi-object Tracking. *Int. J. Comput. Vis.* **2021**, *129*, 548–578. [CrossRef]
43. Suter, R.; Miladinović, D.; Schölkopf, B.; Bauer, S. Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness. In Proceedings of the 36th International Conference on Machine Learning, Long Breach, CA, USA, 9–15 June 2019; Technical Report; IEEE: Piscataway, NJ, USA, 2019. [CrossRef]
44. Black, E.; Wang, Z.; Fredrikson, M.; Datta, A. Consistent counterfactuals for deep models. *arXiv* **2021**, arXiv:2110.03109. [CrossRef]
45. Qian, Y.; Barthelemy, J.; Karuppiah, E.; Perez, P. Identifying Re-identification Challenges: Past, Current and Future Trends. *SN Comput. Sci.* **2024**, *5*, 937. [CrossRef]
46. Alkanat, T.; Bondarev, E.; De With, P. Enabling Open-Set Person Re-Identification for Real-World Scenarios. *J. Image Graph.* **2020**, *8*, 26–36. [CrossRef]
47. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. Mars: A video benchmark for large-scale person re-identification. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11-14 October 2016, Proceedings, Part VI 14*; Springer: Cham, Switzerland, 2016. [CrossRef]
48. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *arXiv* **2016**, arXiv:1609.01775. [CrossRef]
49. 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, Venice, Italy, 22–29 October 2017. [CrossRef]
50. Wu, J.; Huang, Y.; Gao, M.; Gao, Z.; Zhao, J.; Zhang, H.; Zhang, A. A two-stream hybrid convolution-transformer network architecture for clothing-change person re-identification. *IEEE Trans. Multimed.* **2023**, *26*, 5326–5339. [CrossRef]
51. Li, Y.; Lian, G.; Zhang, W.; Ma, G.; Ren, J.; Yang, J. Heterogeneous feature-aware Transformer-CNN coupling network for person re-identification. *PeerJ Comput. Sci.* **2022**, *8*, e1098. [CrossRef]
52. Zhao, S.; Gao, C.; Zhang, J.; Cheng, H.; Han, C.; Jiang, X.; Guo, X.; Zheng, W.S.; Sang, N.; Sun, X. Do Not Disturb Me: Person Re-identification Under the Interference of Other Pedestrians. *arXiv* **2020**, arXiv:2008.06963. [CrossRef]
53. Rao, Y.; Chen, G.; Lu, J.; Zhou, J. Counterfactual attention learning for fine-grained visual categorization and re-identification. *arXiv* **2021**, arXiv:2108.08728. [CrossRef]
54. Sun, Z.; Zhao, F. Counterfactual attention alignment for visible-infrared cross-modality person re-identification. *Pattern Recognit. Lett.* **2023**, *168*, 79–85. [CrossRef]
55. Hirzer, M.; Beleznai, C.; Roth, P.; Bischof, H. Person Re-Identification by Descriptive and Discriminative Classification. In *Image Analysis, Proceedings of the 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, 23–25 May 2011, Proceedings 17*; Springer: Berlin/Heidelberg, Germany, 2011. [CrossRef]

56. Wang, T.; Gong, S.; Zhu, X.; Wang, S. LNCS 8692—Person Re-identification by Video Ranking. In *Computer Vision—ECCV 2014 13th European Conference, Zurich, Switzerland, 6–12 September 2014, Proceedings, Part IV*; Technical Report; Springer: Berlin/Heidelberg, Germany, 2014. [CrossRef]
57. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893. [CrossRef]
58. Cosar, S.; Bellotto, N. Human Re-Identification with a Robot Thermal Camera using Entropy-based Sampling. *J. Intell. Robot. Syst.* **2019**, *98*, 85–102. [CrossRef]
59. Kumar, S.; Yaghoubi, E.; Das, A.; Harish, B.; Proença, H. The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, Re-Identification and Search from Aerial Devices. *arXiv* **2020**, arXiv:2004.02782. [CrossRef]
60. Yin, J.; Wu, A.; Zheng, W. Fine-Grained Person Re-identification. *Int. J. Comput. Vis.* **2020**, *128*, 1654–1672. [CrossRef]
61. Siv, R.; Mancas, M.; Sreng, S.; Chhun, S.; Gosselin, B. People Tracking and Re-Identifying in Distributed Contexts: PoseTReID Framework and Dataset. In Proceedings of the 2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, 6–8 October 2020. [CrossRef]
62. Wang, X.; Paul, S.; Raychaudhuri, D.; Liu, M.; Wang, Y.; Roy-Chowdhury, A. Learning Person Re-identification Models from Videos with Weak Supervision. *arXiv* **2020**, arXiv:2007.10631. [CrossRef]
63. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Liu, S.; Zhou, W. LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. *arXiv* **2021**, arXiv:2108.10831. [CrossRef]
64. Du, Y.; Lei, C.; Zhao, Z.; Dong, Y.; Su, F. Video-Based Visible-Infrared Person Re-Identification With Auxiliary Samples. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 1313–1325. [CrossRef]
65. Zhao, Y.; Shen, X.; Jin, Z.; Lu, H.; Hua, X.S. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; Technical Report; IEEE: Piscataway, NJ, USA, 2020. [CrossRef]
66. Xiang, S.; Fu, Y.; You, G.; Liu, T. Unsupervised Domain Adaptation Through Synthesis For Person Re-Identification. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020. [CrossRef]
67. Zhang, S.; Luo, W.; Cheng, D.; Yang, Q.; Ran, L.; Xing, Y.; Zhang, Y. Cross-Platform Video Person ReID: A New Benchmark Dataset and Adaptation Approach. In *Computer Vision—ECCV 2024 18th European Conference, Milan, Italy, 29 September–4 October 2024, Proceedings, Part XXVII*; Springer: Berlin/Heidelberg, Germany, 2024. [CrossRef]
68. Hambarde, K.A.; Mbongo, N.; MP, P.K.; Mekewad, S.; Fernandes, C.; Silahtaroğlu, G.; Nithya, A.; Wasnik, P.; Rashidunnabi, M.; Samale, P.; et al. DetReIDX: A stress-test dataset for real-world UAV-based person recognition. *arXiv* **2025**, arXiv:2505.04793. [CrossRef]
69. Nguyen, H.; Nguyen, K.; Pemasiri, A.; Liu, F.; Sridharan, S.; Fookes, C. AG-VPReID: A Challenging Large-Scale Benchmark for Aerial-Ground Video-based Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville TN, USA, 11–15 June 2025. [CrossRef]
70. Glymour, C.; Zhang, K.; Spirtes, P. Review of causal discovery methods based on graphical models. *Front. Genet.* **2019**, *10*, 524. [CrossRef]
71. Locatello, F.; Bauer, S.; Lucic, M.; Rätsch, G.; Gelly, S.; Schölkopf, B.; Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Technical Report; IEEE: Piscataway, NJ, USA, 2019. [CrossRef]
72. Cui, Z.; Zhou, J.; Peng, Y.; Zhang, S.; Wang, Y. DCR-ReID: Deep component reconstruction for cloth-changing person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 4415–4428. [CrossRef]
73. Nguyen, V.; Mantini, P.; Shah, S. CrossViT-ReID: Cross-Attention Vision Transformer for Occluded Cloth-Changing Person Re-Identification. In Proceedings of the Asian Conference on Computer Vision (ACCV), Hanoi, Vietnam, 8–12 December 2024. [CrossRef]
74. Liang, T.; Jin, Y.; Gao, Y.; Liu, W.; Feng, S.; Wang, T.; Li, Y. CMTR: Cross-modality transformer for visible-infrared person re-identification. *IEEE Trans. Multimed.* **2021**, *25*, 8432–8444. [CrossRef]
75. Mishra, R.; Mondal, A.; Mathew, J. Nystromformer based cross-modality transformer for visible-infrared person re-identification. *Sci. Rep.* **2025**, *15*, 16224. [CrossRef]
76. Wang, Y.; Zhang, P.; Gao, S.; Geng, X.; Lu, H.; Wang, D. Pyramid Spatial-Temporal Aggregation for Video-based Person Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; Technical Report; IEEE: Piscataway, NJ, USA, 2021. [CrossRef]
77. Wu, P.; Wang, L.; Zhou, S.; Hua, G.; Sun, C. Temporal Correlation Vision Transformer for Video Person Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canasa, 20–27 February 2024; Technical Report; IEEE: Piscataway, NJ, USA, 2024. [CrossRef]
78. Kasantikul, R.; Kusakunniran, W.; Wu, Q.; Wang, Z. Channel-shuffled transformers for cross-modality person re-identification in video. *Sci. Rep.* **2025**, *15*, 15009. [CrossRef] [PubMed]

79. Kocaoglu, M.; Snyder, C.; Dimakis, A.G.; Vishwanath, S. CausalGAN: Learning causal implicit generative models with adversarial training. *arXiv* **2017**, arXiv:1709.02023. [CrossRef]
80. He, T.; Jin, X.; Shen, X.; Huang, J.; Chen, Z.; Hua, X.S. Dense interaction learning for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; Technical Report; IEEE: Piscataway, NJ, USA, 2023. [CrossRef]
81. Chen, Y.; Yang, Y.; Liu, W.; Huang, Y.; Li, J. Pose-guided counterfactual inference for occluded person re-identification. *Image Vis. Comput.* **2022**, *128*, 104587. [CrossRef]
82. Sun, J.; Li, Y.; Chen, L.; Chen, H.; Wang, M. Dualistic Disentangled Meta-Learning Model for Generalizable Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2024**, *20*, 1106–1118. [CrossRef]
83. Kansal, K.; Wong, Y.; Kankanhalli, M. Privacy-Enhancing Person Re-Identification Framework—A Dual-Stage Approach. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; Technical Report; IEEE: Piscataway, NJ, USA, 2024. [CrossRef]
84. Brkljač, B.; Brkljač, M. Person detection and re-identification in open-world settings of retail stores and public spaces. *arXiv* **2025**, arXiv:2505.00772. [CrossRef]
85. Fu, Y.; Wang, X.; Wei, Y.; Huang, T. STA: Spatial-Temporal Attention for Large-Scale Video-based Person Re-Identification. *arXiv* **2018**, arXiv:1811.04129. [CrossRef]
86. Liu, Y.; Yan, J.; Ouyang, W. Quality Aware Network for Set to Set Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Technical Report; IEEE: Piscataway, NJ, USA, 2017. [CrossRef]
87. Chen, D.; Doering, A.; Zhang, S.; Yang, J.; Gall, J.; Schiele, B. Keypoint Message Passing for Video-based Person Re-Identification. *arXiv* **2021**, arXiv:2111.08279. [CrossRef]
88. Alsehaim, A.; Breckon, T. VID-Trans-ReID: Enhanced Video Transformers for Person Re-identification. In Proceedings of the BMVC 2022: The 33rd British Machine Vision Conference, London, UK, 21–24 November 2024; Technical Report; Durham University: Durham, UK, 2022.
89. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014. [CrossRef]
90. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017. [CrossRef]
91. Burgess, C.P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; Lerchner, A. Understanding disentangling in beta-VAE. In Proceedings of the NIPS Workshop on Learning Disentangled Representations, Long Beach, CA, USA, 9 December 2017. [CrossRef]
92. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35. [CrossRef]
93. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial Discriminative Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2021; pp. 7167–7176. [CrossRef]
94. Zhang, Y.; Nie, S.; Liu, W.; Xu, X.; Zhang, D.; Shen, H.T. Sequence-to-Sequence Domain Adaptation Network for Robust Text Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2740–2749. [CrossRef]
95. Wang, Y.; Wang, L.; You, Y.; Zou, X.; Chen, V.; Li, S.; Huang, G.; Hariharan, B.; Weinberger, K. Resource Aware Person Re-identification across Multiple Resolutions. *arXiv* **2018**, arXiv:1805.08805. [CrossRef]
96. Shu, R.; Bui, H.H.; Narui, H.; Ermon, S. A DIRT-T Approach to Unsupervised Domain Adaptation. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
97. Gabbay, A.; Hoshen, Y. Style Generator Inversion for Image Enhancement and Animation. *arXiv* **2019**, arXiv:1906.11880. [CrossRef]
98. Chen, R.T.Q.; Li, X.; Grosse, R.B.; Duvenaud, D.K. Isolating Sources of Disentanglement in Variational Autoencoders. *arXiv* **2018**. Available online: https://arxiv.org/abs/1802.04942 (accessed on 27 June 2025).
99. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein Generative Adversarial Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 214–223. [CrossRef]
100. Lee, H.Y.; Tseng, H.Y.; Huang, J.B.; Singh, M.; Yang, M.H. Diverse Image-to-Image Translation via Disentangled Representations. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018. [CrossRef]
101. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal Unsupervised Image-to-Image Translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018. [CrossRef]

102. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. [CrossRef]
103. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. [CrossRef]
104. Yao, S.; Ardabili, B.; Pazho, A.; Noghre, G.; Neff, C.; Tabkhi, H. Real-World Community-in-the-Loop Smart Video Surveillance—A Case Study at a Community College. *arXiv* **2023**, arXiv:2303.12934. [CrossRef]
105. Gabdullin, N.; Raskovalov, A. Google Coral-based edge computing person reidentification using human parsing combined with analytical method. *arXiv* **2022**, arXiv:2209.11024. [CrossRef]
106. Ghiță, A.; Florea, A. Real-Time People Re-Identification and Tracking for Autonomous Platforms Using a Trajectory Prediction-Based Approach. *Sensors* **2022**, *22*, 5856. [CrossRef]

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