Clothing Purification with Causality Meets Vision-Language Pretraining Models

Zhengwei Yang[1,2†], Huilin Zhu[3†], Nan Lei[4], Basura Fernando[5,6,7], Zheng Wang[1,2*]

[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Wuhan University.
[2]Hubei Key Laboratory of Multimedia and Network Communication Engineering.
[3]School of Computer Science and Artificial Intelligence, Wuhan University of Technology.
[4]School of Computer Science and Electronic Engineering, Hunan University.
[5]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore.
[6]Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore.
[7]College of Computing and Data Science, Nanyang Technological University, Singapore.

*Corresponding author(s). E-mail(s): wangzwhu@whu.edu.cn;
Contributing authors: yzw_aim@whu.edu.cn; jsj_zhl@whut.edu.cn; nancy77@hnu.edu.cn; fernando_basura@cfar.astar.edu.sg;
†These authors contributed equally to this work.

Abstract

Vision-Language Pretraining (VLP) models have shown significant promise in scene understanding and representation learning. However, their application in fine-grained tasks like Cloth-Changing Person Re-Identification (CC-ReID) is challenging due to their reliance on unstable discriminative features such as clothing. Conversely, expert CC-ReID models possess exceptional fine-grained comprehension skills but struggle to obtain reliable cloth-agnostic representations, hindered by the time-consuming and labor-intensive process of obtaining precise annotations and the spurious data associations brought by the co-occurrence phenomenon of identity and clothing. This paper introduces the **Ca**usality-based **Pu**rification (CaPu) model. The CaPu constructs a clothing indication pipeline that leverages the unique strengths of multiple VLP models to efficiently capture clothing semantics. Utilizing these semantics, CaPu employs causality analysis to purify the relationship between learned visual features and intrinsic identity representation from two causal aspects: the Consistency Treatment Effect (CTE) and the Distinctiveness Treatment Effect (DTE). The CTE enhances feature consistency within each identity by simulating clothing changes. Meanwhile, the DTE enhances the model's ability to perceive intrinsic identity representation. Extensive experiments on three standard CC-ReID datasets demonstrate that CaPu achieves state-of-the-art performance.

Keywords: Cloth-changing person re-identification, Vision-language pretraining, Causal inference, Causal intervention.

1 Introduction

The imperative task of identifying individuals across diverse cameras amidst clothing variability presents a formidable challenge in real-world surveillance systems. This highlights the importance of the Cloth-Changing Person Re-IDentification (CC-ReID)

task (Barbosa et al, 2012). The challenge of CC-ReID lies in the notable changes of the target person's clothing over time, compounding the traditional complexities encountered in ReID, such as alterations in viewpoint (Wu et al, 2022, 2023a), instances of occlusion (Somers et al, 2023; He et al, 2023), variations in illumination (Zhang et al, 2022; Lu et al, 2024), and more. Addressing these obstacles requires dedicated research focused on capturing a dependable, fine-grained intrinsic person identity representation, impervious to the variances introduced by clothing.

In the quest for intrinsic person identity representation, existing CC-ReID methods can be categorized into two primary groups: direct and indirect methods. The direct methods explore human body semantics as prior knowledge, *e.g.*, mask (Hong et al, 2021; Li et al, 2023c; Yu et al, 2020; Gao et al, 2022), gait (Jin et al, 2022; Li et al, 2023e), shape (Shi et al, 2022; Chen et al, 2022a; Cui et al, 2023; Qian et al, 2020), and 3D model (Bansal et al, 2022; Chen et al, 2021; Yu et al, 2022a). The direct methods aim to sidestep the influence of clothing on identity representation learning. The representation learning pipeline is defined as:

$$\boldsymbol{F} = E(X, K(X)), \tag{1}$$

where the identity representation $\boldsymbol{F}$ is obtained by encoding the person X and integrating prior knowledge parsing methods K, where E denotes the encoder. However, actuating and integrating prior knowledge requires substantial computational costs. Moreover, by emphasizing certain clothing-irrelevant knowledge, like shape or gait, the direct methods lack a comprehensive understanding of the influence of clothing on identity recognition.

The second category reinforces identity representation through indirect learning-based methods to utilize prior knowledge, *e.g.*, clustering (Liu et al, 2023a), adversarial learning (Yang et al, 2023b), metric learning (Shu et al, 2021b), and contrastive learning (Liu et al, 2023d). Indirect methods concentrate on the mining relationships between features without explicit guidance of body semantics knowledge. The representation learning pipeline is defined as:

$$\boldsymbol{F} = L(E(x_1), E(x_2), \ldots, E(x_n)), \tag{2}$$

where the identity representation $\boldsymbol{F}$ is obtained by conducting learning-based methods L on the images $\{x_1, x_2, \ldots, x_n\} \in X$ after the encoder E. Nevertheless, the absence of explicit prior knowledge guidance

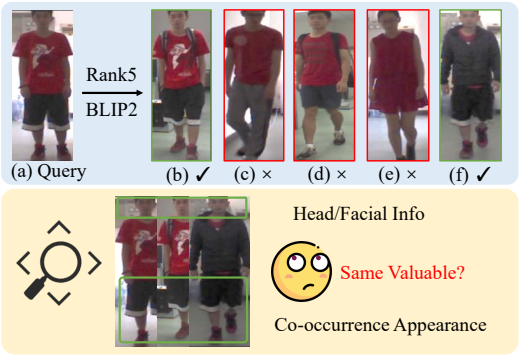
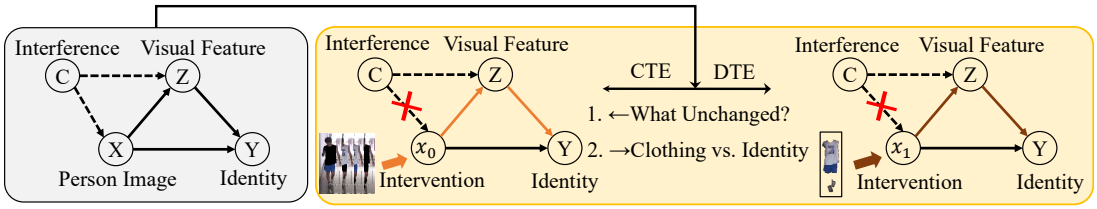


Figure 1 Illustration of the retrieval result by BLIP2 (Li et al, 2023a) and spurious data associations brought by the co-occurrence phenomenon of identity and clothing. The green box denotes the correct retrieval. The red box denotes the incorrect retrieval.

makes them susceptible to potential data association or biases. In summary, both direct and indirect methods encounter challenges in effectively acquiring and utilizing prior knowledge.

To accurately acquire prior knowledge, a model with robust scene understanding and generalization ability is imperative. In recent years, Vision-Language Pretraining (VLP) models, *e.g.*, SAM (Kirillov et al, 2023), GroundingDINO (Liu et al, 2023c), and BLIP2 (Li et al, 2023a) have garnered significant attention. Their remarkable ability of scene understanding and multi-modal comprehension has sparked great excitement in industry and academia. This enthusiasm has led to a rapid surge of groundbreaking applications based on VLP models, which span various domains, including detection (Ming et al, 2022; Du et al, 2022), retrieval (Li et al, 2023d; Chen et al, 2023a; Xie et al, 2023; Siddiqui et al, 2024), and medical diagnoses (Bannur et al, 2023; Zhang et al, 2024), *etc*.

A straightforward insight is to leverage the capability of VLP models to obtain discriminative identity representation. However, as illustrated in Figure 1, even when employing the encoder of BLIP2 (Li et al, 2023a) as a foundational model for CC-ReID, the outcomes still fall prey to clothing interference. The results reveal that general pretraining models have limited proficiency in acquiring discriminative knowledge (Yan et al, 2023; Shao et al, 2023; Wang et al, 2024b), especially in fine-grained tasks like CC-ReID, where individuals frequently change appearance. Precise identity recognition in cloth-changing scenarios requires fine-grained perceptual abilities beyond the capabilities of directly using VLP models. A well-balanced



(a) Causal Graph (b) Intervention & Causality purification

Figure 2 Causal graph for CC-ReID and the illustration of the intervention and causality purification based on this causal graph. The graph denotes the person image X is sent to expert models for visual feature perception Z to recognize identity Y, where the unobserved clothing interference C will affect both X (spurious associations) and Z (confusion attention). x_0 and x_1 are the interventions on the input by modifying clothing and clothing only, where it breaks the connection between C and X.

strategy is needed to leverage VLP's generalization capability while addressing deficiencies in fine-grained perception.

Reevaluating the effectiveness of VLP models within the scope of CC-ReID reveals their limitations in fine-grained perception, but they still have advantages in coarse-grained attribute discovery, such as appearance, shape, and sketch VLP models' general perceptual proficiencies can help alleviate complexities in acquiring prior knowledge. Therefore, the judicious utilization of VLP models provides a straightforward and efficient means to address challenges in the acquisition of prior knowledge, formulated as:

$$K_{\text{vlp}} = \text{VLPs}(X). \tag{3}$$

Acquiring knowledge is a foundational step in research, but effectively leveraging it is vital. Our observations led us to an intriguing phenomenon, prompting our investigation into **how prior knowledge influences identity discrimination.** Theoretically, samples with the same identity but different clothing labels should have distinct outfits. However, the bottom of Figure 1 shows a person retaining some items (*e.g.,* pants, shoes, and the inner shirt) while changing a significant portion of their upper clothing. CC-ReID methods commonly define features that remain unchanged before and after a disguise as "intrinsic identity features". However, the co-occurrence of identity and clothing can introduce spurious associations, confounding the ReID model and leading to inaccurate perceptions of intrinsic identity features. Therefore, breaking spurious data associations is a primary challenge in this paper for utilizing prior knowledge.

Despite the valuable information clothing semantics contribute to human recognition in conventional ReID scenarios, the spurious association between clothing and identity poses a significant challenge in CC-ReID. This association is vulnerable to confusion among inter-class samples with similar appearances, making it difficult to discern the respective contributions of clothing and identity in recognizing target identity. Breaking these spurious associations can be viewed as exploring the causal relationship between learned visual and identity features. Causal inference, a theory that aims to unveil and quantify the causal relation between events or factors, has emerged as a crucial area of research in statistics, epidemiology, and data science. It is widely applied in computer vision tasks such as Visual Question Answering (VQA) (Xue et al, 2023; Li et al, 2023b; Zang et al, 2023; Niu et al, 2021), visual categorization (Rao et al, 2021; Mao et al, 2022; Liu et al, 2022), scene graph generation (Sun et al, 2023; Wu et al, 2023b; Tang et al, 2020), and segmentation (Miao et al, 2023; Ouyang et al, 2022).

In the context of CC-ReID, causal inference provides a powerful methodology to mitigate the spurious associations brought by prior knowledge and uncover the true causal effects underpinning identity recognition. By analyzing data and considering scenarios where a person's clothing changes while their identity remains constant, causal inference gives us an opportunity to purify the visual representation and identify which features are genuinely indicative of identity, a process we term causality purification. This approach enables the development of models that are more robust to changes in clothing and can accurately recognize individuals based on their intrinsic identity features, ultimately improving the performance of CC-ReID systems, which can be defined as:

$$\boldsymbol{F} = E\left(\text{Causal}\left(X, K\left(X\right)\right)\right), \tag{4}$$

$$\boldsymbol{F} = \text{Causal}(E(x_1), E(x_2), \ldots, E(x_n)), \tag{5}$$

for direct and indirect methods, respectively.

In this paper, to overcome the obstacles in acquiring and utilizing prior knowledge, we introduce Causality-based Purification (CaPu) model, which builds upon the successes of VLP and causal inference, defined as:

$$\boldsymbol{F} = E\left(\text{Causal}\left(X, K_{\text{vlp}}\left(X\right)\right)\right), \tag{6}$$

where CaPu integrates a clothing indicator based on VLP models to harness their semantic perceptual capabilities for acquiring prior knowledge and adopting causal analysis on the factors within the image.

Specifically, the clothing indicator in CaPu consists of multiple VLP models, each functioning as a distinct visual comprehension component responsible for attribute discovery, detection, segmentation, and feature extraction, collectively enabling precise clothing indications. Then, to perform causal inference by purifying identity features, we first construct a causal graph, as shown in Figure 2(a), to determine the causal relationships within CC-ReID. Through conducting intervention (Pearl, 2013) on the input, CaPu evaluates the spurious associations between identity and clothing information through two causal perspectives: **C**onsistency **T**reatment **E**ffects (**CTE**) and **D**istinctiveness **T**reatment **E**ffects (**DTE**).

As shown in Figure 2(b), the **CTE** represents the input samples of gradual clothing change, and the effect is how the model perceives the feature. Determining the **CTE** assesses the model's responsiveness to clothing alterations. From a causality standpoint, substantial feature variations due to clothing modifications suggest flawed identity interpretation, while minimal changes likely link to intrinsic identity. Therefore, CaPu selectively applies segmentation masks to clothing items, creating wardrobe groups to break spurious associations between identity and clothing information, which is the first-level causality purification. Furthermore, **DTE** quantifies the change in identity recognition probability when intervening on the samples, particularly when the intervention relates to clothing. Similar identity perception for learned visual and clothing features suggests the model erroneously considers clothing semantics as intrinsic identity. Quantifying and amplifying the distinctiveness between learned visual and clothing features emphasizes the causality between visual representation and intrinsic identity, serving as the second-level causality purification.

In summary, CaPu provides an effective and resource-efficient strategy for handling obstacles in direct and indirect CC-ReID methods, purifying learned identity features for precise CC-ReID, the specific contributions are three-fold:

- We introduce a Causality-based Purification (CaPu) model to provide clothing semantics indications and purify the causality between the learned visual representation and the intrinsic identity representation.
- We establish a Vision-Language Pretraining (VLP) pipeline, which offers a paradigm for integrating VLP models that consider their unique strengths for efficient prior knowledge acquisition.
- We redefine the challenge of breaking spurious associations between identity and clothing information as a causal inference problem. The treatment effects of consistency and distinctiveness are proposed to emphasize intrinsic identity representation for effective knowledge utilization.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related work presented in recent years; Section 3 delves into the proposed method with detailed flowcharts and explanations; Section 4 provides a comprehensive analysis of experiments, including qualitative and quantitative results as well as validation of each module's effectiveness; and finally, we conclude the paper in Section 5.

2 Related Work

2.1 Cloth-changing Person Re-identification

The distinction between traditional person ReID and CC-ReID is that the latter involves individuals disguising their appearances, making clothing information a potential impediment to identity discrimination. This highlights the importance of CC-ReID methods in mining fine-grained clothing-agnostic representation and excavating intrinsic identity information from deep visual semantics. Based on the difference in obtaining clothing-agnostic representation, existing CC-ReID methods can be broadly divided into two categories: direct and indirect.

The direct methods involve exploring human body semantics or heavier patterns, *e.g.*, mask (Hong et al, 2021; Li et al, 2023c; Yu et al, 2020; Gao et al, 2022; Peng et al, 2024), gait (Jin et al, 2022; Li et al, 2023e), shape (Li et al, 2021; Shi et al, 2022; Chen et al, 2022a; Cui et al, 2023; Qian et al, 2020), and 3D model (Bansal et al, 2022; Chen et al, 2021; Yu et al, 2022a), to bypass the interference of clothing in identity representation learning. Specifically, Liu *et al.* (Liu et al, 2023b) leverage masks to separate human parts and assign weights adaptively to

identify challenging regions for comprehensive training. Zhang *et al.* (Zhang et al, 2023a) introduce multi-biological learning, including head, neck, and shoulders, by estimating masks and keypoint to resist cloth-changing. Cui *et al.* (Cui et al, 2023) disentangle clothing-relevant and -agnostic features by reconstructing body contours. Shi *et al.* (Shi et al, 2022) leverage human parsing estimation to enhance the attention of human head part. Jin *et al.* (Jin et al, 2022) and Li *et al.* (Li et al, 2023e) consider gaits as biological features and leverage gait recognition to pursue cloth-agnostic representation. Yu *et al.* (Yu et al, 2022a) and Bansal *et al.* (Bansal et al, 2022) introduce 3D human mesh to obtain multi-modal geometry representations. Ci *et al.* (Ci et al, 2023) and Tang *et al.* (Tang et al, 2023) incorporate multiple tasks and datasets to perform human-centric perception. While incorporating prior knowledge can aid identity recognition, it can also be computationally expensive and not fully account for the influence of clothing. Consequently, the accuracy of prior knowledge and the image quality become a limiting factor in their performance.

The indirect methods reinforce identity representation through data associations and leverage clustering (Han et al, 2023; Li et al, 2022b), adversarial learning (Gu et al, 2022; Yang et al, 2023a), metric learning (Shu et al, 2021b), and contrastive learning (Yang et al, 2022a,b), attention mining (Yang et al, 2023b) to obtain cloth-agnostic representation. Specifically, Liu *et al.* (Li et al, 2022b) propose cloth-aware center cluster loss to gather the intra-class center under different clothing. Gu *et al.* (Gu et al, 2022) adopt adversarial learning by penalizing the model's predictive power on clothing. Shu *et al.* (Shu et al, 2021b) aggregate multiple metrics learning methods to optimize the mAP value during training. Yang *et al.* (Yang et al, 2022a,b) leverage generative adversarial learning by feature cross-wise integration to sample independent feature representation. Additionally, several researchers focus on data augmentation (Kweon and Cho, 2023; Jia et al, 2022) to emphasize cloth-agnostic information. Nevertheless, the absence of explicit prior knowledge guidance in indirect methods leads to an overemphasis on data correlations, making them vulnerable to noisy and inaccurate information.

The common challenge shared between direct and indirect methods centers around the acquisition and utilization of prior knowledge, which constitutes the pivotal challenge within the CC-ReID task.

2.2 Vision-Language Pretraining Model

Recently, numerous Vision-Language Pretraining (VLP) methods exhibit vital ability in scene understanding and representation learning. Their impressive generalizability has led to substantial improvements to various downstream tasks, such as detection (Ming et al, 2022; Du et al, 2022; Song et al, 2022; Dou et al, 2022), retrieval (Li et al, 2023d; He et al, 2023; Chen et al, 2023a; Bai et al, 2023; Xie et al, 2023), and medical diagnoses (Bannur et al, 2023; Zhang et al, 2024; Chen et al, 2022b; Yan and Pei, 2022).

Specifically, CLIP (Radford et al, 2021) extracts the features of vision and language separately and aligns them by contrastive learning. BLIP (Li et al, 2022a) introduces multimodal mixture of encoder-decoder to align multimodal information. BLIP2 (Li et al, 2023a) further combines pertaining models of vision and language to bootstrap multimodal understanding ability. GroudingDINO (Liu et al, 2023c) introduces language to close-set detection to enhance the generalization ability for open-set detection. Segment Anything (SAM) (Kirillov et al, 2023) is built on a promptable segmentation task and supports flexible prompts to facilitate zero-shot performance on numerous tasks. The VLP models' strong multi-modal comprehension and scene understanding capabilities make them excellent foundation models that significantly extend the limits of deep learning.

Several recent efforts have integrated VLP into retrieval tasks, CLIP-ReID (Li et al, 2023d) exploits the cross-modal description ability in CLIP through learnable text tokens for each identity to fine-tune CLIP encoder for ReID. RGANet (He et al, 2023) enhances occluded person ReID by generating human part regions with CLIP and selecting informative regions. UNIReID (Chen et al, 2023a) employs task-specific modality learning to extract and integrate visual and textual information from multiple modalities. RaSa (Bai et al, 2023) leverages VLP models as the backbone to perform unimodal representation learning for text-based person search. MLVR (Xie et al, 2023) combines CLIP to enhance vehicle attribute matching for text-vehicle retrieval.

2.3 Causal Inference in Computer Vision

Causal inference (Kuang et al, 2020; Guo et al, 2020; Pearl, 2013) is concerned with identifying the factual relation between events beyond data association (Pearl, 2010). Causality mining is a popular research area in computer vision that aims to provide insightful and

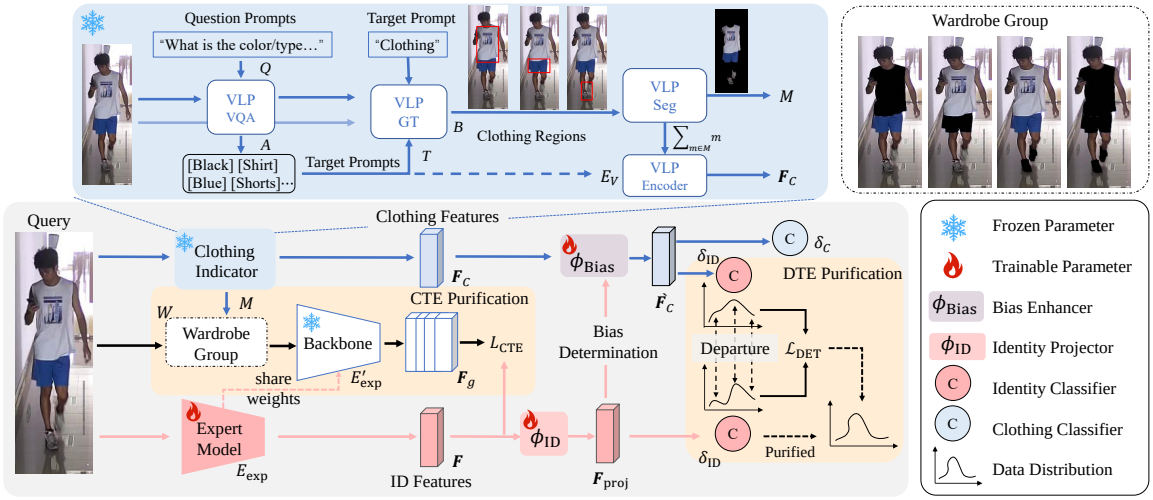


Figure 3 The overall illustration of CaPu. The clothing indicator handles the clothing knowledge accusation by combining the strength of VLP models with frozen parameters. The wardrobe group is constructed based on clothing indications, which tends to break the association between clothing by analyzing the semantic Consistency Treatment Effect. The parameter of the group feature extractor is shared with the expert model and is frozen to maintain the comprehension of targets. After the ID projector, the feature dimensions between clothing branch (blue) and identity feature branch (red) are aligned. Then, the bias enhancer learned to emphasize the biased representation of clothing. Finally, the Distinctiveness Treatment Effect is considered based on the reaction of the identity classifier when facing different inputs for further feature purification.

explainable methods. The concept of intervention and counterfactual (Pearl, 2013; Kuang et al, 2020) is central to causality mining, where the "what if" scenarios play a central role in determining causality. Researchers ask questions like "What would have happened if a different course of action were taken?" and use such insights to assess causality.

The causality mining has been successfully used in several areas, including debiasing (Dash et al, 2022; Liu et al, 2022; Tang et al, 2020), stable learning (Zhang et al, 2021; Liu et al, 2021), and disentanglement learning (Yang et al, 2021a; Yu et al, 2022b). Several methods are proposed within ReID that involve causality mining between the model and data. Yang *et al.* (Yang et al, 2023a) introduce causality into CC-ReID and propose an elaborately designed clothing representation learning branch to distill clothing interference. Similarly, Li *et al.* (Li et al, 2023f) leverage causal intervention to alleviate the strong cloth-identity spurious correlation. Yang *et al.* (Yang and Tian, 2023) and Zhang *et al.* (Zhang et al, 2023b) leverage causal inference to disentangle the correlation between domain and class in Domain Generalization Person Identification (DG-ReID). Liu *et al.* (Li et al, 2022c) use counterfactual inference to construct graph topology structure in visible-infrared ReID.

Beyond the integration of causality with reID, an increasing number of studies are exploring the application of causal methodologies across a wider range of computer vision tasks. VideoQA (Zang et al, 2023) and CF-VQA (Niu et al, 2021) attempt to discover the real association by explicitly capturing visual features that are causally related to the question semantics and weakening the impact of local language semantics on question answering. Chen *et al.* (Chen et al, 2023b), Lv *et al.* (Lv et al, 2022) and MatchDG (Mahajan et al, 2021) infer the causes of domain shift to facilitate domain generalization ability in DG-ReID. CAL (Rao et al, 2021) employs random attention to formulate counterfactual causality for visual categorization. The emergence of causal inference enables considering assumptions about reality that were previously unattainable. This provides us with theoretical support to break spurious data associations and purify the learned visual representation.

3 Method

The main structure of the proposed Causality-based Purification (CaPu) model is illustrated in Figure 3. CaPu comprises a clothing indicator and a causality-based feature purification module. We begin by formulating the knowledge-acquiring process within the clothing indicator in Section 3.2 followed by a detailed description of causality-based purification by computing the intervention probability $P(Y|do(X))$ in Section 3.3. The specific prompt design for each step in the clothing indicator is shown in Section 3.4.

3.1 Problem Formulation

For a given image of person x^i with identity label y^i, the image is first fed into the clothing indicator to obtain masks of clothing item $M^i = \{m_1^i, m_2^i, \dots, m_k^i\}$, where k stands for the number of clothing items, all clothing masks are denoted as $m_A^i = \sum m^i$. Based on person image x^i and clothing masks M^i, we can systematically generate a set of images to build a wardrobe group for each person. The wardrobe group $W^i = \{w_1^i, w_2^i, \dots, w_k^i, w_A^i\}$, depicts the gradual change in clothing items of a given person, w_A^i stands for all the clothing items are masked. Simultaneously, pure clothing masks m_A^i are sent to VLP encoder E_V to obtain clothing representation $\boldsymbol{F}_C^i$. The person image x^i is sent to expert model E_{exp} to obtain person representation $\boldsymbol{F}^i$. At the training stage, the expert model is trainable but the parameters within the clothing indicator are fixed. Only the purified expert model with feature $\boldsymbol{F}$ is used at the testing stage.

3.2 Clothing Indicator

As illustrated in the blue region in Figure 3, the clothing indicator consists of several Vision-Language Pretraining (VLP) models to capture clothing prior knowledge. The clothing indicator is built on strong cross-modal comprehension and coarse-grain representation capabilities of VLP models, including Visual Question Answering (VQA), Grounding, and Segmentation models. These models work together to generate masks for clothing items and extract meaningful clothing representations Each visual comprehension component within the Clothing Indicator has distinct responsibilities, including attribute discovery, detection, and segmentation. This collaborative effort ensures accurate and detailed clothing indications. The Clothing Indicator utilizes natural language question prompts $Q = \{q_1, q_2, \dots\}$ to control precise understanding of VLP models to visual entities, which consists of several steps to obtain clothing representation. Each step of the prior knowledge acquisition process is formulated as:

$$A = \{a_1, a_2, \dots\} = \text{VQA}(q_1, q_2, \dots), \tag{7}$$

where question prompts q are sent to VQA models, resulting in answer set A with clothing attributes. Subsequently, these answers are selected and reconstructed based on the clothing states, forming the target prompt set T. Then T is sent to grounding model GT for generating bonding box B of each target, defined as:

$$T = \{t_1, t_2, \dots\} = \text{Select}(A), \tag{8}$$

$$B = \{b_1, b_2, \dots\} = \text{GT}(T), \tag{9}$$

then, these bounding boxes with attributes are then fed into the Segmentation model Seg to generate clothing masks accordingly, defined as:

$$M = \{m_1, m_2, \dots\} = \text{Seg}(B), \tag{10}$$

finally, the masks for individual clothing items are combined, representing the clothing, and are sent to a VLP encoder to generate the overall clothing representation, defined as:

$$\boldsymbol{F}_C = E_V(\sum_{m \in M} m). \tag{11}$$

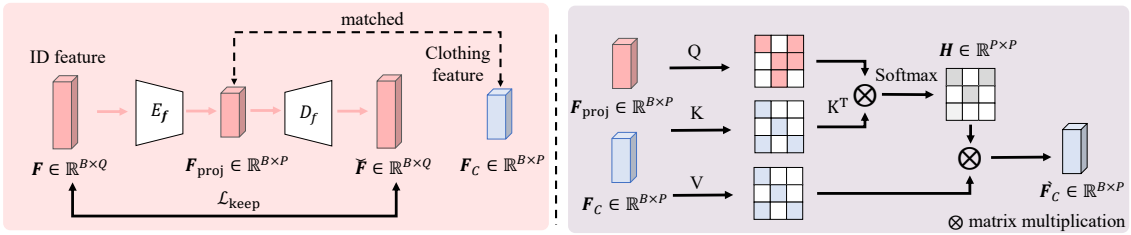
Through the synergy of various VLP models, we forge a comprehensive automated knowledge acquisition pipeline. This pipeline streamlines the accurate localization of clothing regions and enables precise extraction of clothing representations. The rich clothing-related prior knowledge gathered through this pipeline lays a solid foundation for the ensuing process of identity feature purification by the CaPu model. For further insights into the specific prompt designs for each component, please refer to Section 3.4.

3.3 Causality Purification

Through harnessing clothing semantics and feature-based prior knowledge, CaPu then transforms the challenge of mitigating the impact of spurious associations between intrinsic identity and noisy data into a problem of causal inference. Specifically, by digging into the relation between clothing and identity, CaPu assesses spurious associations between identity and clothing information from two causal perspectives: Consistency Treatment Effects and Distinctiveness Treatment Effects.

3.3.1 Consistent Treatment Effects

Determining the **Consistent Treatment Effects (CTE)** relies on assessing the model's responsiveness to alterations in clothing. From a causality standpoint, if modifications in clothing lead to substantial variations in features, it suggests a flawed model interpretation of identity. Conversely, features demonstrating minimal changes amid clothing alterations are more likely to be linked with intrinsic identity. Therefore, CaPu seeks



(a) Identity Projector (b) Bias Enhancer

Figure 4 The detailed structure of the Identity Projector and Bias Enhancer. Specifically, (a) the Identity Projector utilizes an encoder-decoder structure to project learned identity features, while preserving as much information from the original features as possible and reducing information loss during projection. (b) the Bias Enhancer amplifies the impact of clothing on identity recognition by combining frozen clothing features with trainable identity features. This interaction allows for the dynamic acquisition of clothing information that affects identity recognition.

to discern causal relations by examining how alterations in one variable (*e.g.*, clothing) influence another (*e.g.*, identity feature). Such intervention on clothing representation is formulated as:

$$P(Y|do(X)) = \sum_{w} P(Y|X, W = w)P(W = w), \quad (12)$$

where do(X) is do-calculate (Fenton et al, 2020), X and Y are separately denote the model input and output, w is the sample in wardrobe group W.

To put this intervention into practice, CaPu initially obtains the visual feature from the expert model upon receiving the image of a person. Concurrently, to preserve the model's comprehension of identity and reduce the additional data noise contributed by the background, we duplicated and then froze the current parameter of the encoder as E'_{exp} to obtain features corresponding to the different clothing states of the individual within the wardrobe group W. Following Eq.(12), we can obtain the identity features group $\boldsymbol{F}_g = \{\boldsymbol{F}_{g_1}, \boldsymbol{F}_{g_2}, \ldots, \boldsymbol{F}_{g_k}, \boldsymbol{F}_{g_A}\}$. Subsequently, the consistent treatment effect can be determined by measuring the difference of identity features $\boldsymbol{F}$ with group identity features $\boldsymbol{F}_g$. And the model's consistency can be achieved by minimizing the $P(Y|do(X))$. To maximize the model's consistency and minimize interference from similar backgrounds between samples, we introduce $\mathcal{L}_{\text{CTE}}$ followed by triplet loss to minimize the dissimilarity between $\boldsymbol{F}$ and $\boldsymbol{F}_g$, formulated as:

$$\mathcal{L}_{\text{CTE}} = \sum_{z=1}^{k+1} \max(d_z^- - d_z^+ + \alpha, 0), \quad (13)$$

$$d_z^+ = min(d(\boldsymbol{F}, \boldsymbol{F}_{g_z}^+)), \quad d_z^- = max(d(\boldsymbol{F}, \boldsymbol{F}_{g_z}^-)), \quad (14)$$

where d_z^+ stands for the distance of the hardest positive pair between visual feature $\boldsymbol{F}$ and the z-th group feature $\boldsymbol{F}_{g_z}^+$ with same identity, and d_z^- denotes the distance of the hardest negative pair between visual feature $\boldsymbol{F}$ and the z-th group feature $\boldsymbol{F}_{g_z}^-$ with different identities. $k+1$ denotes the number of samples in the wardrobe group for each identity. α is a marginal hyperparameter. Building upon reinforcing CTE, the $\mathcal{L}_{\text{CTE}}$ aims to emphasize features invariant to clothing variations, and it encourages the model to learn feature representations that exhibit greater consistency under changing clothing conditions. Additionally, measuring the distance between samples with different identities allows the mitigation of common interferences such as background noise. Therefore, CTE helps improve the model's ability to recognize identities regardless of clothing changes and background variations, which is the first-level feature purification.

While the concept of the wardrobe group may appear intuitive, its design is grounded in a causal framework that guides the structured separation of identity- and clothing-specific features. By leveraging pedestrian attributes and carefully designed partial masking, this structure helps preserve body part semantics critical to disentangling confounding factors. Unlike previous methods such as body jigsaw (Zhou et al, 2022) or direct cloth-changing approaches (Shu et al, 2021a; Shi et al, 2022), which often rely on explicit masking or generation-based manipulation, our method reduces sensitivity to image quality and clothing state. This robustness is achieved through a powerful VLP pipeline that enables more reliable segmentation, and a CTE module that focuses on intrinsic identity information even when facial or clothing features are absent or altered.

The CTE encapsulates the core idea of using causal reasoning to address feature entanglement, emphasizing the principles behind feature purification rather than being tied to a specific implementation. The focus on causal relationships between identity and clothing features provides a structured foundation that

enables learning in a more interpretable and theoretically robust manner. This perspective extends beyond individual technical solutions, offering a guiding principle for future advancements in the disentanglement of identity-related and clothing-related information.

3.3.2 Distinctiveness Treatment Effects

Relying solely on CTE can only constrain the feature at the surface level, which does not consider the profound interference of clothing to identity. To mitigate the diverse impact of spurious associations, clothing representations are employed to further regulate deep-level representations. To diminish the expert model's identity recognition based on clothing, CaPu evaluates the response of identity classifier to different features to measure **Distinctiveness Treatment Effects (DTE)**. To achieve this, two obstacles are unavoidable in CaPu: feature alignment and bias determination.

Identity Projector. After obtaining the features for clothing and the person separately, a challenge arises due to the inherent misalignment between the feature spaces generated by the VLP encoder and the expert model. This misalignment can be attributed to structural disparities; the VLP encoder predominantly employs a Transformer architecture, in contrast to the Convolutional Neural Network (CNN) commonly applied in the expert models. The former prioritizes global relations and lacks translation invariance, while the latter focuses on local information with more flexible adaptability to change. Therefore, we focus on the identity features of the expert model for alignment to minimize information loss from VLP features. To achieve this, we propose an Identity Projector that aligns the features from the expert model with those from the VLP encoder.

As shown in Figure 4(a), to achieve feature alignment while retaining as much original information as possible, CaPu incorporates an encoder-decoder framework. The visual feature obtained by expert model $\boldsymbol{F} \in \mathbb{R}^{B\times Q}$ undergoes downgrading by passing through the encoder E_f, producing $\boldsymbol{F}_{\text{proj}} \in \mathbb{R}^{B\times P}$, where B stands for batchsize, Q is the final feature dimension of the identity feature, P is the final feature dimension of the clothing feature. Then, $\boldsymbol{F}_{\text{proj}}$ is reconstructed through the decoder D_f to obtain $\breve{\boldsymbol{F}} \in \mathbb{R}^{B\times Q}$, which is the same dimension of $\boldsymbol{F}$. Concurrently, an information maintaining constraint $\mathcal{L}_{\text{keep}}$ is enforced to ensure that the features pre- and post-encoding-decoding converge toward consistency. The identity projector aligns the features of different branches while preserving information before and after projection, formulated as:

$$\boldsymbol{F}_{\text{proj}} = E_f(\boldsymbol{F}) \qquad \breve{\boldsymbol{F}} = D_f(\boldsymbol{F}_{\text{proj}}), \tag{15}$$

$$\mathcal{L}_{\text{keep}} = ||\boldsymbol{F} - \breve{\boldsymbol{F}}||_2, \tag{16}$$

where $\boldsymbol{F}_{\text{proj}}$ has the same dimension as clothing feature $\boldsymbol{F}_C$.

Bias Enhancer. Considering the parameters of the clothing indicator to be frozen, the extracted clothing features remain fixed. To leverage these fixed clothing representations to guide the learnable identity features and to explicitly delineate the impact of clothing on identity features, CaPu introduces a bias enhancer. As shown in Figure 4(b), this bias enhancer explicitly identifies the portions of the identity features influenced by clothing. Specifically, cross-attention mechanisms is employed to dynamically capture the interactions between clothing and identity features, formulated as:

$$\boldsymbol{H}(\boldsymbol{F}_{\text{proj}}, \boldsymbol{F}_C) = \text{softmax}(\frac{\boldsymbol{F}_{\text{proj}}\boldsymbol{F}_C^T}{\sqrt{d_C}}), \tag{17}$$

$$\grave{\boldsymbol{F}}_C = \boldsymbol{H}(\boldsymbol{F}_{\text{proj}}, \boldsymbol{F}_C)\boldsymbol{F}_C, \tag{18}$$

where $\boldsymbol{H}$ is a weight matrix, the softmax function is applied to compute the attention scores for each feature pair. d_C is a scaling factor with the same dimension as $\boldsymbol{F}_C$. Additionally, a clothing discriminator is introduced to regulate the feature attention to clothing, ensuring that clothing-related information is retained in the fused representations, which is constrained by a cross-entropy loss $\mathcal{L}_C$ among the clothing discriminator outputs, formulated as:

$$\mathcal{L}_C = -\sum_{i=1}^{N} \log \frac{e^{(\delta_C(\grave{\boldsymbol{F}}_C^i)/\tau)}}{\sum_{j=1}^{N_C} e^{(\delta_C(\grave{\boldsymbol{F}}_C^j)/\tau)}}, \tag{19}$$

where N is the number of samples. N_C is the number of suits, calculated as the total number of independent suits associated with each identity. δ_C denotes the clothing discriminator. The temperature parameter $\tau \in \mathcal{R}^+$ is a hyper-parameter to control the output scale of classifier, thereby controlling the distribution of class decisions.

So far, we gathered all the elements required to analyze DTE, where the treatment refers to inputs of clothing and identity features, while the effect is the probability of the model recognizing identities. The DTE seeks to quantify the change in the probability distribution of identity recognition when the feature space

is intervened upon, specifically, when the intervention is related to clothing features.

$$P_{\text{ID}} = P(Y|do(X = \boldsymbol{F}_{\text{proj}})), \tag{20}$$

$$P_C = P(Y|do(X = \grave{\boldsymbol{F}}_C)), \tag{21}$$

$$\mathcal{L}_{\text{DTE}} = \text{argmax}(\Delta(P_{\text{ID}}, P_C)), \tag{22}$$

where X denotes the learned identity features, and Y stands for the identity recognition results, which have slightly different meanings within Eq. (12). Formally, the DTE is expressed as the maximal distribution distance Δ between two probability distributions: P_{ID}, the identity perception distribution when conducting intervention from the learned visual representation, and P_C, the identity perception distribution when conducting intervention from clothing representation.

To implement DTE, CaPu frames it as a distance departure problem in latent space, involving both orientation and distance for vectors. Considering direct manipulation of features may compromise the integrity of the feature space, CaPu employs an identity classifier constrained by a commonly used identification loss on $\boldsymbol{F}_{\text{proj}}$ to analyze distributions of identity and clothing-influenced features. Specifically, Eq. (20) and Eq. (21) can be implement as:

$$\mathcal{P}_{\text{ID}} = \delta_{\text{ID}}(\boldsymbol{F}_{\text{proj}}), \quad \mathcal{P}_C = \hat{\delta_{\text{ID}}}(\grave{\boldsymbol{F}}_C), \tag{23}$$

where δ_{ID} is the identity classifier and $\hat{\delta_{\text{ID}}}$ denotes the identity classifier shared weights with δ_{ID}, and $\mathcal{P}_{\text{ID}}$ and $\mathcal{P}_C$ are distribution matrices that represent the identity predictions.

Then, the distinction in orientation among two distributions can be quantified by cosine similarity $\boldsymbol{S}$ for each sample in the batch, formulated as:

$$\boldsymbol{S} = \frac{\mathcal{P}_{\text{ID}}\mathcal{P}_C^{\text{T}}}{||\mathcal{P}_{\text{ID}}||\,||\mathcal{P}_C||}. \tag{24}$$

To quantify the spatial distance between distributions in the latent space and accentuate dissimilarities within the same identity context, CaPu introduces a Positive-Reinforced Distance, where samples with the same identity in a batch are served as positive, and the distributions of $\mathcal{P}_{\text{ID}}$ and $\mathcal{P}_C$ within positive samples are encouraged to be departure. This mechanism penalizes the classifier for being too lenient in distinguishing between different instances of the same identity, prompting it to refine its understanding of what constitutes identity-specific features. The distance can be expressed as:

$$\boldsymbol{D} = ||\mathcal{P}_{\text{ID}} - \mathcal{P}_C||^2, \tag{25}$$

$$\boldsymbol{D}' = \sqrt{\beta\left(\sum \boldsymbol{D} \odot \boldsymbol{V} - \sum \boldsymbol{D} \odot (1 - \boldsymbol{V})\right)}, \tag{26}$$

where $\boldsymbol{V}$ denotes the one-hot matrix where each row is a one-hot vector with the index of label y_i corresponding to 1 and others are 0, and $\odot$ is the element-wise multiplication and the sum is over all samples in the batch, β serves as a scaling factor to regulate the sensitivity to the distance between positive samples.

Finally, the DTE in Eq. (22) is implemented as:

$$\mathcal{L}_{\text{DTE}} = \underbrace{(1 + \boldsymbol{S})}_{\text{Orientation}} + \underbrace{max(m - \boldsymbol{D}^{'}, 0)}_{\text{Distance}}, \tag{27}$$

where, to be noted, although $\boldsymbol{S}$ and $\boldsymbol{D}^{'}$ are vectors, they are computed across the batch, with each element representing the cosine similarity and the positive-reinforced distance for each sample between the identity and clothing features. The first term constrains the orientation between two distributions, encouraging an increase in their dissimilarity, with the value being 0 when the dissimilarity between the two distributions is at its maximum. The second term constrains the spatial distance in feature space and does not increase their similarity beyond a certain margin m to prevent the model from diverging.

Maximizing the distance in Eq. (22) can be transferred to minimize $\mathcal{L}_{\text{DTE}}$, which contributes to the purification of identity features. The collaboration of both terms ensures that the distance of two distributions remains bounded, contributing to the second-level feature purification, allowing CaPu to focus on the intrinsic identity representation while minimizing the impact of clothing-related information on identity recognition.

Through a dual-level purification based on causality analysis, CaPu disentangles spurious associations between identity and clothing information, ultimately achieving identity feature purification.

3.4 Prompts Design

In this section, the specific prompt design of each step in the clothing indicator is elaborated upon in detail.

Step1: Clothing Attribute Discovery We devise question-based prompts to fully leverage the advantages of VLP models and meticulously guide their

attention toward clothing items. To achieve this, we employ BLIP2 (Li et al, 2023a) as a basis for Visual Question Answering (VQA) to elucidate and refine clothing components and attributes. This clothing attribute discovery aims to provide a solid foundation for the VLP models to better comprehend clothing and obtain accurate clothing representations. To avoid the potential bias and illusion problems within current VLP models, instead of using prompts to get the general attributes within one round, we delicately developed 8 distinct questions that are tailored to cover a wide range of scenarios related to various clothing states of individuals, as shown below. These comprehensive prompts are intended to ensure that the VLP models can efficiently capture the appearance of the target clothing under different conditions.

Question Prompts:
INPUT:
1. What color of clothing is the person wearing on the upper body?
2. What type of clothing is the person wearing on the upper body?
3. What color of clothing is the person wearing on the lower body?
4. What type of clothing is the person wearing on the lower body?
5. Is the person in the image wearing something on the feet?
6. What color of shoes is the person wearing on the feet if any?
7. What type of shoes is the person wearing on the feet if any?
8. Is the person wearing a dress?
OUTPUT:
Answers based on the given image.

Among the above prompts, questions 1-4 serve as foundational prompts, collectively encompassing most of the regular scenarios. Questions 5-7 are supplementary prompts designed to address instances of missing information about footwear due to suboptimal image quality or angles. Question 8 is an assist prompt to cover the common clothing type found in females. Notably, since there is an overlap in clothing attributes between questions 1-4 and 8, if the answer to question 8 is yes, the results of questions 2 and 4 are discarded.

To be noticed, the inclusion of fine-grained attributes such as color and type in the VLP pipeline was initially motivated by the goal of enhancing flexibility and reusability beyond the immediate scope of this work. While our framework primarily focuses on accurate clothing segmentation, these attributes were intended to support broader annotation needs, such as fine-grained retrieval and attribute-based recognition. However, we acknowledge that not all prompts are essential for the current task. In particular, color-related prompts are not directly used in the downstream segmentation or learning process. As such, the prompt design in our pipeline is modular and can be adapted based on practical requirements. In our tests, removing the color-related questions (reducing the total from 8 to 5) results in approximately a 40% reduction in the GPU usage for the VQA step, offering a meaningful improvement in efficiency without compromising segmentation performance.

Step2: Clothing Region Determine Open-set detection (Liu et al, 2023c) and segmentation (Kirillov et al, 2023) have inspired interesting applications and demonstrated remarkable ability. We show that combining such comprehension ability aids in accurately localizing clothing regions. Specifically, given the attribute prompts, we apply Grounding DINO (Liu et al, 2023c), which is an open-set detection model, to detect the corresponding clothing for each person.

Broadly, each suit of clothing encompasses three fundamental components: the upper body, lower body, and footwear. Besides, an alternative situation exists where the clothing comprises only dress and footwear. In response to these different scenarios and guided by the response from Question 8 in Step 1, the target prompt supplied to the DINO model is dynamically adjusted. Furthermore, to address the potential limitations in the VLP model's sensitivity to fine-grained attributes, we additionally introduced an overarching target prompt, "clothing".

Target Prompts:
INPUT:
Example 1: [clothing, shirt, shorts, shoe, (specific type of shoe)]
Example 2: [clothing, dress, shoe, (specific type of shoe)]
OUTPUT:
Visual regions of clothing items.

Step3: Clothing Mask Generation Given the clothing region prompt in the former step, we leverage the open-set segmentation model SAM (Kirillov et al, 2023) to extract precise masks for each identified clothing item.

Mask Prompts:
INPUT: Visual region prompts from Step 2
OUTPUT: Masks of clothing items.

Using these clothing masks, we generate a set of modified images for each individual, composing a wardrobe group where clothing items are progressively masked (*i.e.*, replaced with black regions) while preserving the rest of the visual content. This black-masking strategy, inspired by prior occlusion-based regularization methods (Singh and Lee, 2017; Huang et al, 2018; Song et al, 2018) provides a neutral suppression of clothing regions, ensuring the model does not associate specific clothing cues. This progressive masking strategy is crucial for disentangling clothing-related features from intrinsic identity cues, as it ensures that

the identity-relevant regions, such as body shape and facial features, remain intact for feature extraction.

Step4: Clothing Representation Extraction Upon obtaining clothing masks, we also aim to capitalize on the expressive potential of VLP models. To achieve this, we extract clothing representations using the encoder of the VLP model with fixed parameters. This strategy capitalizes on the VLP model's inherent general comprehension capabilities to interpret the superficial semantics of clothing. Notably, this method offers a more expedient alternative in contrast to extant supervised feature extraction techniques.

Visual Prompts:
INPUT: Masks of all clothing from Step 3
OUTPUT: Clothing representation.

4 Experiment

4.1 Datasets and Evaluation

4.1.1 Datasets

We comprehensively evaluate the efficacy of CaPu across three publicly available benchmark datasets designed explicitly for CC-ReID. Three datasets include PRCC (Yang et al, 2021b), LTCC (Qian et al, 2020), and VC-Cloth (Wan et al, 2020), each contributing distinct challenges and scenarios for the evaluation of CC-ReID models.

PRCC (Yang et al, 2021b) is an extended-duration indoor person ReID dataset, characterized by three stationary cameras. It encompasses 221 distinct identities and comprises a total of 33,698 images. Notably, individuals in PRCC wear same clothing when captured by cameras A and B in different rooms, while wearing different clothing when captured by cameras C at different times. The training set comprises 150 identities with 17,896 images, whereas the testing set involves the remaining 71 identities, totaling 15,802 images.

LTCC (Qian et al, 2020) is a demanding dataset designed for prolonged surveillance scenarios, spanning days and months. LTCC features challenges such as frequent clothing changes and serious vagueness, which involves 17,119 labeled images of 152 persons under twelve cameras. LTCC is bifurcated into two subsets: a cloth-changing set that contains 14,756 images of 91 persons with 417 different sets of clothing; and a cloth-consistent subset consisting of 2,382 images of the remaining 61 identities without clothing changes.

VC-Cloth (Wan et al, 2020) is a virtual dataset created by an action-adventure game engine GTA5. It consists of 19,060 labeled images of virtual people belonging to 512 different IDs. The dataset includes images captured from four cameras. Cameras 2 and 3 maintain consistent clothing, while there are four separate scenes with varying illumination conditions. For training purposes, there are 256 IDs with a total of 9,449 images. The remaining set for testing comprises another set of 256 IDs with a total of 9,601 images.

4.1.2 Evaluation Settings

We follow previous studies (Gu et al, 2022; Yang et al, 2023a; Rao et al, 2021) and leverage the standard rank at K (R@K) accuracy and mean average precision (mAP) for evaluation. R@K is defined as the fraction of queries where the correct items are among the top-K gallery items, and we use the strictest criterion of $K = 1$ for evaluation. Following the baseline (Gu et al, 2022), we perform experiments in two experimental settings: standard and changing settings. The former includes both samples with changed clothing and samples with consistent clothing, while the latter only includes samples with changed clothing. It is important to mention that for PRCC and VC-Cloth datasets, the standard (SC) setting only includes clothing-consistent samples. In the standard setting, images with the same ID and camera view in the testing set are excluded from evaluation. In the changing setting, besides the same ID and camera view, samples with the same clothing are also excluded during testing to evaluate the model's performance on unseen clothing.

Additionally in PRCC, there are two strategies for evaluating under the cloth-changing setting: single- and multi-shot matching. The single-shot setting randomly selects a different outfit image for each ID as a gallery for testing, creating a small gallery for evaluation. The multi-shot setting includes all images with both cloth-changing and cloth-consistent elements for testing, resulting in a large gallery. While some methods use single-shot matching multiple times and average the results, the multi-shot setting is widely preferred because it reduces result fluctuations caused by randomness. To ensure a fair comparison, we report the results under the multi-shot setting in Table 1 to compare with state-of-the-art methods, while marking the results under single-shot as $*$.

Table 1 Comparison of R@K (%) and mAP (%) performance with the state-of-the-arts on small scale datasets. "†" denotes the methods that are designed for CC-ReID. "‡" indicates the reproduced results. "*" represents the results from a small gallery. **Bold** numbers are the best results. Prior "-" means the indirect method using mathematical calculations to introduce prior knowledge. All notations remain consistent throughout the following.

Method	Venue	Prior	PRCC				LTCC				VC-Cloth					
			Standard (SC)		Changing		Standard		Changing		Standard		Standard (SC)		Changing	
			R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP
PCB (Sun et al, 2018)	ECCV 18	-	99.8	97.0	41.8	38.7	65.1	30.6	23.5	10.0	87.7	74.6	94.7	94.3	62.0	62.2
OSNet (Zhou et al, 2019)	ICCV 19	-	-	-	-	-	67.9	32.1	23.9	10.8	-	-	-	-	-	-
HPM (Fu et al, 2019)	AAAI 19	-	99.4	96.9	40.4	37.2	66.7	31.6	24.5	10.9	86.8	77.2	94.2	94.0	65.9	64.0
IANet (Hou et al, 2019)	CVPR 19	-	99.4	98.3	46.3	46.9	63.7	31.0	25.0	12.6	-	-	-	-	-	-
ABD-Net (Chen et al, 2019)	ICCV 19	-	-	-	49.2	44.8	-	-	-	-	87.1	81.0	-	-	80.9	79.2
ISP (Zhu et al, 2020)	ECCV 20	Mask	92.9	-	36.9	-	66.7	30.2	28.8	12.4	89.1	83.6	95.0	94.5	80.1	74.9
AGW (Ye et al, 2021)	TPAMI 21	-	98.8	91.7	38.9	34.2	68.8	36.5	32.9	13.8	92.2	85.4	94.9	94.5	80.4	75.2
TransReID (He et al, 2021)	ICCV 21	-	97.5	96.1	47.7	50.2	72.8	38.3	31.9	17.1	90.9	81.2	95.1	94.6	71.7	72.5
RCSANet (Huang et al, 2021) †	ICCV 21	-	99.6	96.6	48.6	50.2	-	-	-	-	-	-	-	-	-	-
3DSL (Chen et al, 2021) †	CVPR 21	Shape	-	-	51.3	-	-	-	31.2	14.8	-	-	-	-	79.9	81.2
FSAM (Hong et al, 2021) †	CVPR 21	Mask	98.8	-	54.5*	-	73.2	35.4	38.5	16.2	-	-	94.7	94.8	78.6	78.9
GI-ReID (Jin et al, 2022) †	CVPR 22	Gait	79.0	-	33.3	-	63.2	29.4	23.7	10.4	-	-	-	-	64.5	57.8
UCAD (Yan et al, 2022) †	IJCAI 22	Mask	96.5	-	45.3	-	74.4	34.8	32.5	15.1	-	-	-	-	-	-
ViT-VIBE (Bansal et al, 2022) †	WACV 22	Shape	99.7	-	47.0	-	71.4	35.8	-	-	-	-	-	-	-	-
IRANet (Shi et al, 2022) †	IVC 22	Pose	99.7	97.8	54.9	53.0	-	-	-	-	-	-	-	-	-	-
Pos-Neg (Jia et al, 2022) †	TIP 22	Shape	-	-	54.8*	-	75.6	37.0	36.2	14.4	-	-	-	-	-	-
ACID (Yang et al, 2023b)†	TIP 23	-	99.1	99.0	55.4*	66.1*	65.1	30.6	29.1	14.5	-	-	95.1	94.7	84.3	74.2
MBUNet (Zhang et al, 2023a)†	TIP 23	Pose	99.8	99.6	68.7*	65.2*	67.6	34.8	40.3	15.0	-	-	95.4	95.3	82.7	70.3
DCR-ReID (Cui et al, 2023)†	TCSVT 23	Mask	**100.0**	99.7	57.2	57.4	76.1	42.3	41.1	20.4	-	-	-	-	-	-
AIM (Yang et al, 2023a)†	CVPR 23	-	**100.0**	**99.9**	**57.9**	**58.3**	76.3	41.1	40.6	19.1	-	-	-	-	-	-
CVSL (Nguyen et al, 2024)†	WACV'24	Pose	97.5	99.1	57.5	56.9	76.4	41.9	**44.5**	**21.3**	-	-	-	-	-	-
Baseline (Gu et al, 2022)‡	CVPR 22	-	**100.0**	99.7	54.4	54.2	73.4	39.4	38.0	17.4	92.3	86.7	95.0	95.2	81.9	78.9
CaPu (Ours)		VLM-Mask	**100.0**	**99.9**	**57.9**	57.2	**76.5**	**43.1**	41.3	20.5	**93.8**	**89.6**	**95.6**	**95.6**	**88.0**	**83.9**

4.1.3 Implementation Details

The expert model is a ResNet50 (He et al, 2016) trained by CAL (Gu et al, 2022), where the last pooling layer and the fully connected layer are removed. Following the detailed settings in previous methods (Gu et al, 2022; Qian et al, 2020; Hong et al, 2021), random horizontal flipping, random cropping, and random erasing (Zhong et al, 2020) are used for data augmentation. CaPu is trained with batchsize of 32. The Adam (Kingma and Ba, 2015) optimizer is adopted in CaPu. Specifically, the parameter of the expert model is fixed at the first 25 epochs for increasing stability, and the initial learning rate is $3.5e{-}6$ and drops to 10% of the original at the 30th epochs. The initial learning rate of the other parts is $3.5e{-}5$, which drops to 10% of the original for every 20 epochs. The temperature parameter τ in Eq. (19) is set to 1/16. The margin α in Eq. (12) and m in Eq. (27) is set to 0.3, and the scaling factor β in Eq. (26) is set to 16. All hyperparameters are fixed for all datasets without further tuning. All classifiers mentioned in the paper are implemented as a single fully connected layer.

4.2 Comparison with State-of-the-Art Methods

In Table 1, we conduct a comprehensive evaluation of the proposed CaPu against eight traditional ReID methods and fourteen methods specifically designed for CC-ReID across three datasets. To ensure a fair and consistent comparison, we replicate the results of the baseline method (Gu et al, 2022) using officially published codes and subsequently employ these weights for training CaPu. Given the limited scope of available comparative methodologies, CaPu is benchmarked against both direct and indirect methods. To highlight the distinctive characteristics of these two approaches, an additional column labeled "prior" is incorporated into Table 1. This column signifies the potent prior knowledge imposed by direct category methods.

As illustrated in Table 1, the proposed CaPu demonstrates a notable superiority over most methodologies and exhibits a substantial performance improvement over the baseline method (Gu et al, 2022) across all evaluated metrics and datasets. This establishes CaPu as the current state-of-the-art in CC-ReID, showcasing its remarkable efficacy in addressing the challenges posed by clothing-changing scenarios. Specifically,

in the indoor PRCC dataset, characterized by minimal pedestrian variations and limited environmental changes in the standard setting, most methods, including CaPu, achieve commendable performance. However, in the Changing setting, CaPu and AIM (Yang et al, 2023a) emerge as the top-performing methods, achieving a competitive R@1 accuracy of 57.9%. Notably, CaPu exhibits a slightly higher mAP than AIM in this context. This discrepancy is attributed to AIM, like CaPu, leveraging causal relations between clothing and identity for identity feature learning. Nevertheless, AIM also employs clothing features without strong supervisory information, leading to a potential reduction in precision concerning clothing features. Consequently, AIM and CaPu yield comparable results in scenarios with clear data and minimal environmental interference like PRCC. In contrast, on more complex datasets LTCC, CaPu outperforms AIM. This highlights the superiority of CaPu, combining VLP and causal reasoning for feature purification, demonstrating its effectiveness over fixed dual-branch methods.

For the real-world scenarios in LTCC, CaPu achieves 76.5%/43.1% on R@1/mAP in the standard setting, surpassing the baseline method (Gu et al, 2022) by 3.1%/3.7%. In the clothing-changing setting, CaPu achieves competitive results, with 41.3%/20.5% on R@1/mAP, surpassing the baseline by 3.3%/3.1% and ranking as the second-best method overall. The slightly lower performance in this setting compared to CVSL (Nguyen et al, 2024) reflects the complementary nature of the two approaches. CVSL, which leverages human keypoints and graph attention networks to distill shape-related features, is particularly well-suited for clothing-changing scenarios. However, its reliance on skeletal information makes it sensitive to factors such as clothing fit (e.g., loose garments) and environmental conditions, which can affect retrieval accuracy. In contrast, CaPu's focus on purifying identity-specific features ensures robust performance across both settings, demonstrating its versatility and effectiveness in mitigating spurious associations between identity and clothing.

Furthermore, among all the competitors, the proposed CaPu achieves the best performance on the VC-Cloth dataset, where CaPu achieves the highest 88.0%/83.9% on R@1/mAP in the changing setting, surpassing others by a large margin of 3.7%/2.7% at least. Besides, CaPu outperforms other methods on the rest settings, achieving 93.8%/89.6% and 95.6%/95.6% on R@1/mAP in two standard settings, respectively. The above results clearly suggest the benefit of CaPu on VC-Cloth by combining the VLP models and the expert model for feature purification, showing the superiority of CaPu in handling synthetic samples.

Table 2 Ablation studies of each component in CaPU under the cloth-changing setting. The black dot in solid indicates that the specific module is considered at the training stage.

Basic		Causal		PRCC		LTCC	
B/L	Group	$\mathcal{L}_{\text{CTE}}$	$\mathcal{L}_{\text{DTE}}$	R@1	mAP	R@1	mAP
●	○	○	○	54.4	54.2	38.0	17.4
●	○	●	○	54.4	54.3	38.2	17.7
●	○	○	●	56.7	56.9	39.7	19.3
●	●	●	○	56.8	56.5	40.8	19.6
●	●	●	●	**57.9**	**57.2**	**41.3**	**20.5**

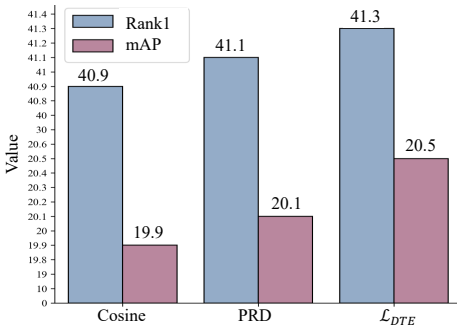


Figure 5 Ablation studies of the two components in $\mathcal{L}_{\text{DTE}}$, "Cosine" denotes the cosine similarity (Orientation), and "PDR" stands for the positive-reinforced distance (Distance) in the DTE loss

4.3 Ablation Studies

To explore the effectiveness of the proposed modules in CaPu, we conduct comprehensive ablation studies from multiple perspectives.

4.3.1 Analysis on Components

The Effectiveness of Each Component. In Table 2, we conducted ablation experiments on each component of CaPu. "B/L" is abbreviated for baseline, and "Group" denotes the constructed wardrobe group. The two losses in the "Causal" section represent the two parts of the treatment effect. Solid circles indicate the participation of the component in the training process, while hollow circles indicate the opposite. It's noteworthy that there is a connection between "Group" and $\mathcal{L}_{\text{CTE}}$ where "Group" serves as the foundation for $\mathcal{L}_{\text{CTE}}$. Considering only $\mathcal{L}_{\text{CTE}}$ without "Group" implies using the image itself as the content for the wardrobe, reducing $\mathcal{L}_{\text{CTE}}$ to a triplet loss.

Specifically, the first row denotes the results of the baseline method. The comparison of rows 1-3 reveals

Table 3 Comparing different designs of the wardrobe group.

Method	PRCC		LTCC	
	R@1	mAP	R@1	mAP
Baseline	54.4	54.2	38.0	17.4
CaPu w/ Direct Group	53.1	53.0	36.7	17.0
CaPu w/ DGNet (Zheng et al, 2019)	54.3	54.9	37.7	17.3
CaPu w/ SCHP (Li et al, 2020)	56.5	56.2	39.1	18.4
CaPu w/ VLP (ours)	**57.9**	**57.2**	**41.3**	**20.5**

the impact of applying the investigation of the two treatment effects without the wardrobe group. Adding the simple version of $\mathcal{L}_{\text{CTE}}$ for training shows only a slight improvement, but with the addition of DTE, there is a significant enhancement of 2.7% and 1.9% in mAP on two datasets, demonstrating the effectiveness of DTE for feature purification. By comparing the results of rows 2 and 4, it can be affirmed that the design of the wardrobe group effectively enhances model performance, indicating the positive significance of CTE for feature purification. Lastly, by comparing the results of the last row with all previous rows, it can be inferred that each component of CaPu has positively contributed to the final outcome, confirming the effectiveness of each component.

Moreover, the $\mathcal{L}_{\text{DTE}}$ encompasses two pivotal components designed to optimize the distribution when intervening with identity and clothing representations. To elucidate the distinct impact of each component, Figure 5 delineates the model performance changes when considering different terms. From the figure, we can infer that while individually applying each strategy facilitates distribution separation, their combined effect leads to a more pronounced distribution separation and consequently yields better results.

The Effectiveness of VLP Pipeline. One major innovation in CaPu lies in constructing a VLP annotation pipeline, harnessing the advantages of VLP models to adeptly capture clothing information for knowledge acquisition. To validate the superiority of the VLP pipeline, we conduct a comprehensive evaluation against alternative methods for constructing the wardrobe group, including SCHP (Li et al, 2020), a prevalent human parser method applied in DCR-ReID (Cui et al, 2023), FSAM (Hong et al, 2021), SPL (Shu et al, 2021a), as well as simpler grouping strategies and GAN-based approaches, on two realistic datasets, as delineated in Table 3.

The "Direct Group" means directly grouping features of the same identity with different clothing labels. This design was chosen as a straightforward baseline because it mimics the process of grouping images within the same identity while assuming sufficient variation in clothing. However, this approach fails to account for the critical challenge of co-occurrence appearance in CC-ReID datasets, where different samples of the same identity often share overlapping clothing attributes (e.g., adding or removing a jacket while retaining the same underlying outfit). This overlap can mislead the model into associating identity-specific features with clothing patterns, thereby degrading its ability to disentangle intrinsic identity features. As shown in Table 3, this approach leads to significantly degraded performance, underscoring the necessity of a more robust method for preserving body part semantics.

The GAN-based approach, using DG-Net (Zheng et al, 2019), was evaluated as a potentially cost-effective alternative for generating diverse clothing variations within the wardrobe group. This method transfers clothing across identities while preserving pose and generates new samples that simulate clothing changes. While this approach introduces variability, it is limited by the inherent instability and artifacts in GAN-generated images. These issues include incomplete clothing changes (e.g., partial adjustments or artifacts) and the loss of critical discriminative details, such as facial features. Moreover, the training of GAN models is dataset-specific, and the process requires significant computational resources and time (close to 24 hours in our experiments), which limits their practicality for real-world CC-ReID tasks. As seen in Table 3, the results obtained using GAN-based methods are only marginally better than the simpler grouping strategy, further highlighting their limitations.

In comparison, SCHP (Li et al, 2020) demonstrates improved results in simpler datasets, such as PRCC, but struggles under complex conditions, such as those in LTCC, where accurate segmentation is crucial. The VLP outperforms all alternative methods by ensuring high-quality annotations and preserving body part semantics, thereby enabling effective disentanglement of identity features from clothing variations.

The design of these experiments reflects our intent to validate the importance of preserving body part semantics and demonstrate the limitations of simpler alternatives. These results highlight the critical role of the VLP pipeline in achieving robust and consistent performance for CC-ReID tasks.

The Effectiveness of Identity Projector. CaPu integrates both clothing and identity features, facing a common challenge of misalignment due to differences in feature dimensions between the VLP models and the

Table 4 Ablation studies of different projector designs in CaPu. "Base" stands for the decoded feature is used for testing. "Linear" refers to direct linear projection between two feature dimensions. "Weight" denotes normalized learnable weight projection. "MLP" stands for the Multilayer Perception Union. "w/o" means not considering the corresponding component in training.

Identity Projector						PRCC		LTCC	
Base	Linear	Weight	MLP	w/o $\mathcal{L}_{\text{keep}}$	Ours	R@1	mAP	R@1	mAP
●	○	○	○	○	○	54.3	54.3	37.8	17.4
○	●	○	○	○	○	56.7	55.8	38.8	18.9
○	○	●	○	○	○	56.1	55.9	40.8	19.1
○	○	○	●	○	○	55.2	55.3	38.0	19.2
○	○	○	○	●	○	57.1	56.6	40.8	19.7
○	○	○	○	○	●	**57.9**	**57.2**	**41.3**	**20.5**

expert models. To address this, the identity projector is strategically employed. Its purpose is to amalgamate the distinctive attributes of both feature spaces, seeking to minimize information loss while maximizing the preservation of the original feature information. This strategic approach ensures optimal feature alignment. The comparative analysis in Table 4 contrasts the proposed identity projector with several conventional feature alignment methodologies.

The results in row 1 demonstrates that features processed through the identity projector retain their performance compared to the original identity features, indicating only a minor information shift attributable to the inherent randomness in neural networks. Upon scrutiny of rows 2-4 and the final row, it becomes evident that direct feature mapping (Weight, Leaner, MLP) for alignment yields only incremental improvements in performance. This underscores the limitation of direct mapping methodologies, emphasizing their deficiency in imposing constraints on information consistency. This, in turn, results in notable information loss and compromises the precision of feature alignment, thereby impacting overall performance. In a noteworthy validation of the efficacy of information consistency constraints, row 4 demonstrates results obtained without leveraging consistency constraint loss $\mathcal{L}_{\text{keep}}$. The discernible contrast with the final row reinforces the positive impact of information consistency on facilitating superior feature alignment. This nuanced constraint proves pivotal in fortifying alignment accuracy, allowing CaPu to adeptly execute the intricate task of feature purification.

The Effectiveness of Bias Enhancer. The purpose of the Bias Enhancer is to establish a dynamic correspondence between the frozen clothing representation and the trainable identity representation, emphasizing the effect of clothing bias within identity features. In contrast to the direct utilization of all clothing features, the Bias Enhancer is designed to capture clothing information influencing identity discernment, facilitating CaPu in enhancing the precision of feature purification. Table 5 illustrates two direct methodologies employing all clothing information, *i.e.*, Linear, and MLP, compared with CaPu's Bias Enhancer.

Table 5 Ablation studies of different bias enhancer designs in CaPu. "Linear" refers to direct linear projection between two feature dimensions. "MLP" stands for the Multilayer Perception Union. Reverse means the features representing Q and K, V in Figure 4(b) are reversed, where the clothing feature is treated as Q.

Bias Enhancer				PRCC		LTCC	
Linear	MLP	Reverse	Ours	R@1	mAP	R@1	mAP
●	○	○	○	57.7	57.0	40.3	19.2
○	●	○	○	56.4	56.1	41.2	19.5
○	○	●	○	57.2	56.7	40.6	19.6
○	○	○	●	**57.9**	**57.2**	**41.3**	**20.5**

Table 6 Comparing different designs of the $\mathcal{L}_{\text{DTE}}$, where w/o $\mathcal{L}_{\text{DTE}}$ stands for the results without consider the DTE part. The NegKL and Marginal NegKL stand for two KL settings without and with margin, respectively.

Method	PRCC		LTCC	
	R@1	mAP	R@1	mAP
Baseline	54.4	54.2	38.0	17.4
w/o $\mathcal{L}_{\text{DTE}}$	56.8	56.5	40.8	19.6
w/ NegKL	57.1	56.6	40.9	19.9
w/ Marginal NegKL	57.0	56.7	40.9	20.0
ours	**57.9**	**57.2**	**41.3**	**20.5**

Through a comparison of rows 1-2 with the final row, it is evident that CaPu's bias-enhancing design exhibits considerable advantages over methodologies relying on the entire clothing features for guidance. This substantiates the efficacy of the Bias Enhancer. Additionally, to validate the effectiveness of using identity features as a query within the enhancer, row 3 presents an experiment using clothing features as query features. The results indicate that even when clothing features are utilized as queries, CaPu maintains a certain level of feature purification. However, with the identity feature as the computation focus, there is an introduction of identity information interference into bias features, affecting the final feature purification outcome. In summary, the design of the Bias Enhancer in CaPu effectively reinforces the expression of clothing bias, promoting the accuracy of feature purification.

4.3.2 Analysis on the alternatives of DTE Loss

The Disentanglement Treatment Effect ($\mathcal{L}_{\text{DTE}}$) is a critical component of our framework, designed to separate identity-specific and clothing-specific features

effectively. To validate its impact, we conducted ablation experiments comparing the full model with and without $\mathcal{L}_{\text{DTE}}$. As shown in Table 6, removing $\mathcal{L}_{\text{DTE}}$ results in a significant performance drop across both PRCC and LTCC datasets, confirming its importance in disentangling these two feature spaces.

To investigate alternative approaches, we explored the use of negative KL divergence as a potential design for $\mathcal{L}_{\text{DTE}}$. KL divergence is a widely used metric to quantify the divergence between two probability distributions, and applying a negative KL divergence can effectively maximize the separation between identity-specific and clothing-specific feature distributions. Given its simplicity and broad applicability, KL divergence represents a natural candidate for this task.

We evaluated two variations: direct negative KL divergence and a margin-based negative KL divergence that introduces a boundary constraint to limit excessive separation. The results of these experiments, shown in Table 6, indicate that both approaches yield slight improvements over the baseline without $\mathcal{L}_{\text{DTE}}$. Direct negative KL divergence achieves marginal performance gains, while the margin-based variation produces comparable results, suggesting that the boundary constraint does not significantly impact performance.

While these results demonstrate the utility of KL divergence in encouraging separation between distributions, neither variation achieves the performance of $\mathcal{L}_{\text{DTE}}$. These findings validate the effectiveness of the proposed $\mathcal{L}_{\text{DTE}}$ in disentangling identity and clothing features, highlighting its role as a key component of our framework.

4.3.3 Analysis on Complexity

To validate CaPu in acquiring and utilizing prior knowledge, we conducted a comparison in Table 7 with several methods in terms of model parameters, FLOPs, and inference time. ISP (Zhu et al, 2020) and AGW (Ye et al, 2021) represent conventional pedestrian re-identification methods, CAL (Gu et al, 2022) is an indirect-type CC-ReID method, FSAM (Hong et al, 2021) is a direct-type CC-ReID method using masks as identity aids, and AIM (Yang et al, 2023a) is a causality based CC-ReID method, exploring the causal relation between clothing and identity features similar to CaPu. It is important to note that due to limited access to open-source implementations, for a fair comparison, CaPu can only rely on existing results provided by their origin papers and has to adopt the single-shot matching strategy on PRCC accordingly, which means the results are produced on a small gallery.

Table 7 Comparisons on the model parameters (Params), FLOPs, training and testing time.

Method	Training			Testing			PRCC*
	Params	FLOPs	Time	Params	FLOPs	Time	R@1
ISP	31.7M	-	16.5h	31.7M	-	30s	38.9
AGW	23.8M	262.7G	3.2h	23.8M	260.9G	116s	49.6
CAL	24.1M	262.2G	2.1h	23.5M	262.1G	58s	54.6
FSAM	164.3M	-	12h	23.8M	-	15s	54.5
AIM	143.9M	540.4G	4.1h	23.5M	262.1G	58s	57.8
CaPu (ours)	66.5M	262.7G	2.8 h	23.5M	262.1G	58s	**59.4**

Table 8 Comparisons on the pre-processing computational cost.

Method	GPU usage				Speed
	VQA	Grounding	Segment	Feature	Time (/image)
SCHP (Li et al, 2020)	-	-	1.8G	-	0.2s
VLP (ours)	2.9G	2G	5.5G	1.6G	1.4s

From the results in Table 7, it is observed that, compared to CAL (Gu et al, 2022), although CaPu increases the model parameters by a modest amount (24.1M for CAL vs. 66.5M for CaPu), it incurs only a marginal increase in training time (2.1h vs. 2.8h) while maintaining the same overhead and speed during testing, achieving a significant improvement of 4.8% in performance. Compared to FASM (Hong et al, 2021), CaPu achieves a 4.9% performance improvement with smaller parameters (164.3M vs. 66.5M) and shorter training time (12h vs. 2.8h). In comparison to the causal-based method AIM (Yang et al, 2023a), CaPu attains a 1.6% performance improvement with approximately 50% fewer parameters and training time, demonstrating the superiority of CaPu in effective acquisition and utilization of prior knowledge.

To address the computational cost associated with the VLP pipeline, we conducted a detailed comparison of GPU usage and processing speed with SCHP (Li et al, 2020) in Table 8. Although the SCHP (Li et al, 2020) requires minimal GPU usage, solely for segmentation, and achieves a significantly faster processing time of 0.2 seconds per image. In contrast, the VLP pipeline involves multiple stages, including VQA, grounding, segmentation, and feature extraction, resulting in higher GPU usage and a longer processing time of 1.4 seconds per image. While this increased complexity incurs greater computational costs, it enables the VLP pipeline to achieve significantly better segmentation quality, as demonstrated in Figure 6.

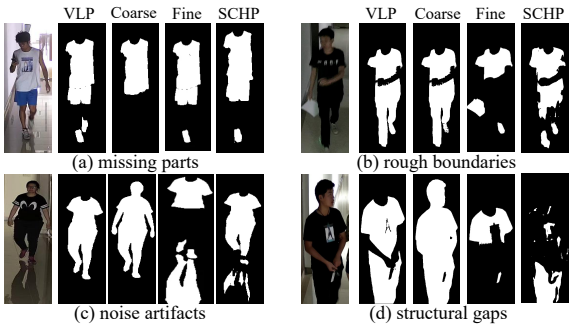


Figure 6 Comparison of segmentation results across several groups: Each group displays RGB images followed by segmentation results from VLP (ours), Coarse (Grounded SAM (Ren et al, 2024) using "clothing" as a prompt), Fine (using "upper cloth", "lower cloth" and "shoes" as a prompt) and SCHP (Li et al, 2020). The comparison highlights the advantages of VLP in addressing issues such as missing parts, rough boundaries, noise artifacts, and structural gaps.

The comparison highlights the trade-off between time complexity and performance. SCHP's simplicity and lower resource consumption make it a viable choice for tasks where computational efficiency is prioritized. However, for scenarios demanding high-quality and robust segmentation, the VLP pipeline proves advantageous. Its ability to address challenges such as missing parts, rough boundaries, and noise artifacts ensures reliable feature extraction, which is critical for downstream tasks requiring fine-grained segmentation.

Additionally, with the rapid development of vision-language models, stronger VQA alternatives such as MiniGPT4 (Zhu et al, 2024), MiniCPM-V (Yao et al, 2024), and Qwen2-VL (Wang et al, 2024a) have become available. These models often provide enhanced visual reasoning capabilities but typically require substantially more GPU memory (often exceeding 7 GB even under quantization). While such models may offer advantages in well-resourced environments, we adopt BLIP-2 in our framework as a practical and efficient choice. With a modest memory footprint (2.9 GB for VQA) and stable performance across diverse input conditions, BLIP-2 offers a good balance between effectiveness and computational efficiency—making it especially suitable in scenarios where accessibility and deployment scalability are key considerations.

4.4 Qualitative Results

4.4.1 Effectiveness of Clothing Indicator

To vividly illustrate the effectiveness of the Clothing Indicator and the clothing perception capability of VLP pipeline, Figure 7 showcases the results for two clothing scenarios. The indication of clothing attributes originates from the results of Indication Step 1 in Section 3.4, while the clothing mask comes from Step 3. Since visualizing other steps exhibits redundancy with Figure7, only partial step results are displayed here. Additionally, based on the clothing indication, the composition of the wardrobe group for the corresponding samples is presented. The results in Figure 7 demonstrate that for typical clothing structures like Top-Bottom-Shoes, the Clothing Indicator in CaPu adeptly perceives clothing edges and accurately identifies clothing components. Similarly, for a different structure like Dress-Shoes, it precisely recognizes the clothing region. Notably, in the second case, the model correctly identifies the clothing region even when the person's hand is present in that area, showcasing CaPu's resilience to such complexities.

To further evaluate its effectiveness, we conducted a detailed comparison against SCHP (Li et al, 2020), a commonly used human parsing model, as well as a coarse segmentation method based on Grounded SAM (Ren et al, 2024) using both general and fine-grained prompts. The results of this comparison are visualized in Figure 6. The results show that the VLP pipeline significantly outperforms alternative methods in handling challenging segmentation scenarios. While it may seem intuitive that more specific prompts (e.g., "upper cloth," "lower cloth," and "shoes") would yield better results, our observations show that they often lead to unstable segmentation. Specifically, fine-grained prompts can introduce structural noise and fragmented outputs, especially under occlusion or non-frontal views. We attribute this to the training bias of the grounding model, which is more reliably aligned with frequently seen, general concepts such as "clothing." In contrast, less common fine-grained terms receive weaker model attention and may even result in over-segmentation or misalignment.

Compared to SCHP and both prompt variants of the SAM-based method, the VLP pipeline achieves more complete, coherent, and noise-free segmentation. It captures subtle and occluded clothing regions with clear boundaries and preserves structural integrity—avoiding the disjointed or noisy masks often observed in other approaches. These results support the advantage of combining VQA-guided grounding with SAM in a modular and reliable way.

Related to the computational cost in Table 8, while the VLP pipeline involves higher computational costs due to the use of multiple models (e.g., VQA, grounding, segmentation, and feature extraction), its superior

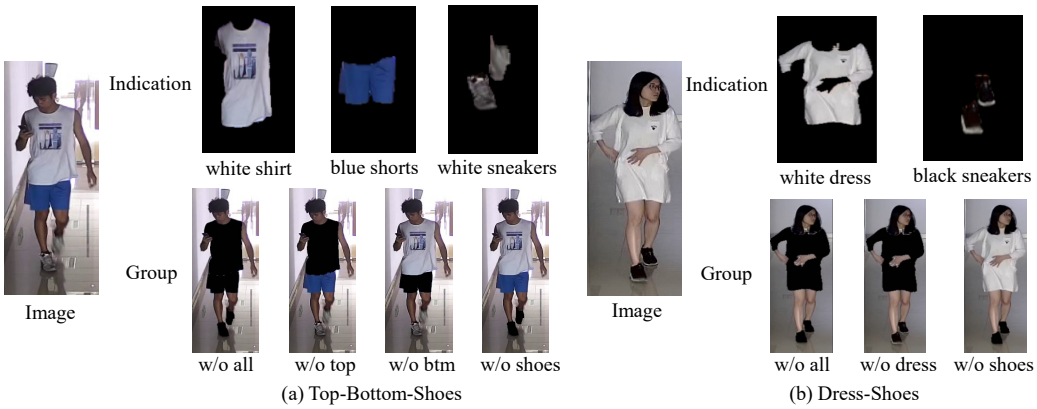


Figure 7 Visualization of the clothing indications and images in the corresponding wardrobe groups of two kinds of standard clothing status.

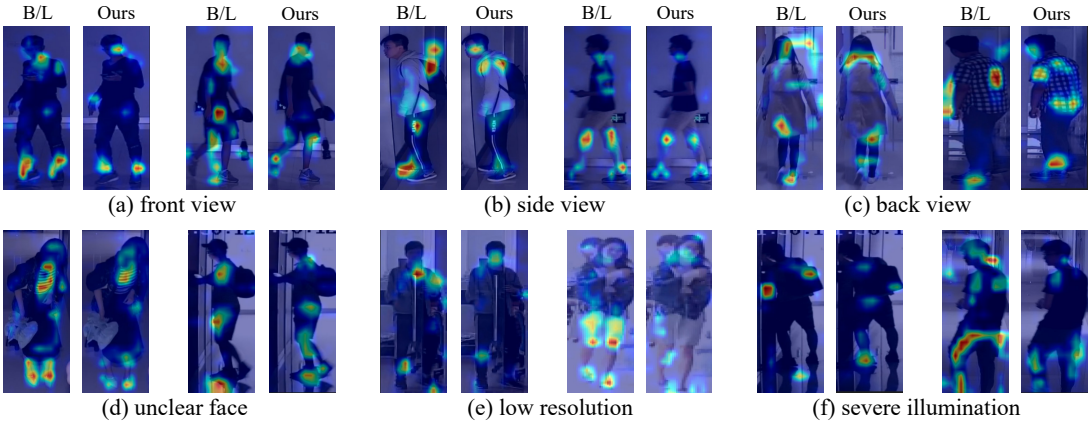


Figure 8 Visualizations of the feature of interest as heatmaps, produced by Grad-CAM (Selvaraju et al, 2020) under various scenarios. Each group corresponds to a specific scenario, with two pairs of samples shown. In each pair, the left heatmap represents the baseline method (labeled as "B/L"), and the right heatmap shows the results of the proposed CaPu method.

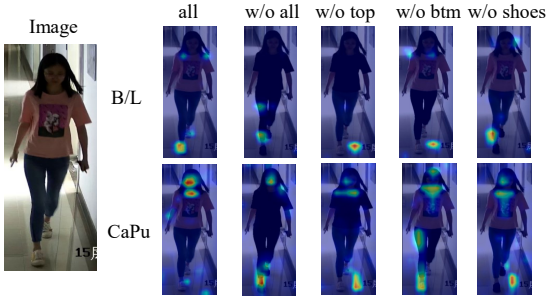


Figure 9 Visualization of the feature of interest as heatmaps, produced by Grad-CAM (Selvaraju et al, 2020). Samples of the same identity but different clothing statuses in the wardrobe group are illustrated. The above row shows the baseline results; the bottom row shows the results of CaPu.

segmentation quality justifies this trade-off. By addressing issues such as missing parts, rough boundaries, and noise artifacts, the VLP pipeline ensures robust and reliable performance, underscoring the importance of incorporating advanced vision-language pretraining in our framework. And the above results confirm that the Clothing Indicator effectively handles the acquisition of prior knowledge.

4.4.2 Effectiveness of Feature Purification

Analysis of Feature Heatmaps In this section, we visually assess the effectiveness of feature purification implemented by CaPu. Figures. 8 and 9 provide attention maps before and after applying feature purification. The attention maps depict the model's focus on specific regions within the images, thereby offering insights into the model's learning dynamics.

Specifically, to evaluate the robustness of CaPu in capturing identity-related features, we conducted additional experiments under various challenging scenarios,

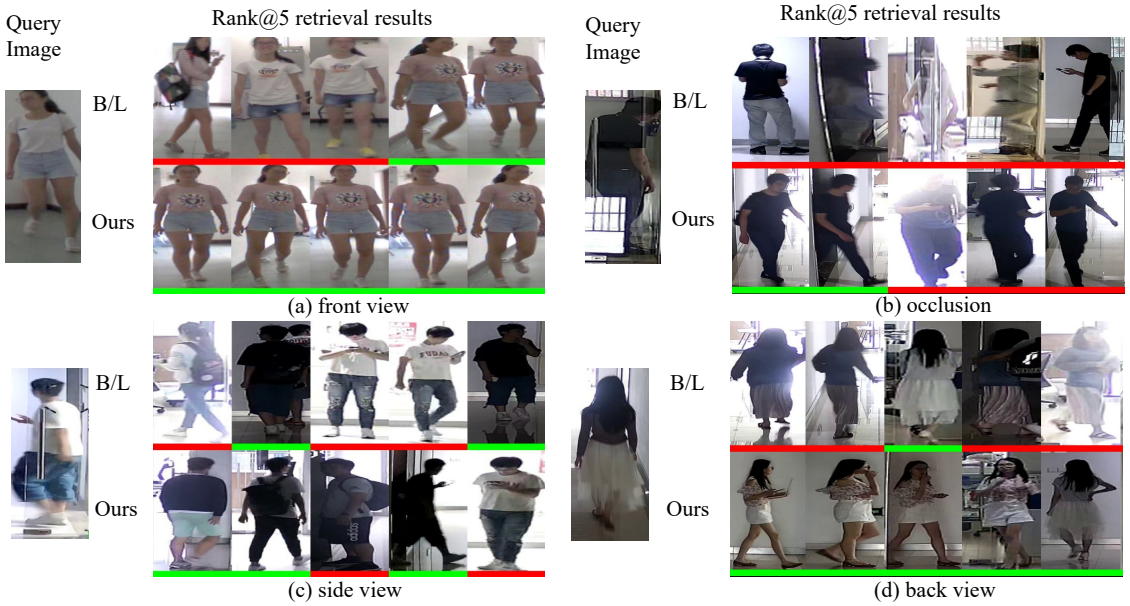


Figure 10 The comparison of retrieval results in various real scenarios. Within each group, the top row shows the baseline results; the bottom row shows the results of CaPu. Green bars denote the correct retrievals, whereas red bars represent the incorrect ones.

as visualized in 8. These Grad-CAM heatmaps compare the baseline method with CaPu across conditions such as occlusions, pose variations, and challenging environmental factors like poor lighting and low resolution. The results demonstrate that the proposed CaPu effectively purifies identity-related features not onIJCVly under ideal frontal-view conditions but also when facial information is insufficient. By focusing on body parts such as the torso and limbs, CaPu ensures robust feature purification and shows strong adaptability across diverse scenarios, enhancing the model's effectiveness in capturing consistent and discriminative identity features.

Furthermore, Figure 9 illustrates the model's attention shift on images from the wardrobe group, highlighting how the model's focus varies. The first column of Figure 9 illustrates the attention map derived from the original RGB image. Notably, the baseline method primarily concentrates its learned features around the feet, neglecting critical intrinsic identity content. This behavior arises due to the model's vulnerability to spurious associations between unchanged clothing parts and intrinsic identity information. The model erroneously prioritizes clothing features as easily learnable identity cues, which compromises accuracy. When analyzing the impact within the wardrobe group, attention maps in columns 2-4 display the effects of selectively removing clothing items from original images. The results in the first row clearly show that despite changes in clothing, the baseline model continues focusing on the foot region. This persistence highlights how clothing information interferes with and introduces bias into the model's representations.

In contrast, CaPu's features demonstrate greater adaptability when faced with changing clothing and encompass identity relevant information. This adaptability reduces bias and confirms the effectiveness of CaPu's feature purification. Notably, when examining attention visualizations, CaPu's features do not always focus on identity regions like the head but correctly concentrate on semantically meaningful regions such as the shoulders, wrists, and ankles. This nuanced focus reflects CaPu's deliberate wardrobe design, which avoids direct masking of clothing sections and instead preserves semantically relevant clues associated with identity. The results underscore CaPu's ability to maintain semantic consistency during feature purification while mitigating spurious associations introduced by clothing variations.

Analysis of Retrieval Results To comprehensively evaluate the feature purification effectiveness of CaPu, we present retrieval results on the real dataset in Figure 10 and the toy dataset in Figure 11.

Specifically, Figure 10 provides retrieval results in real-world scenarios, offering a direct comparison between the baseline method and CaPu under diverse conditions, including frontal view, occlusion, side view, and back view. Juxtaposing the two sets of retrieval

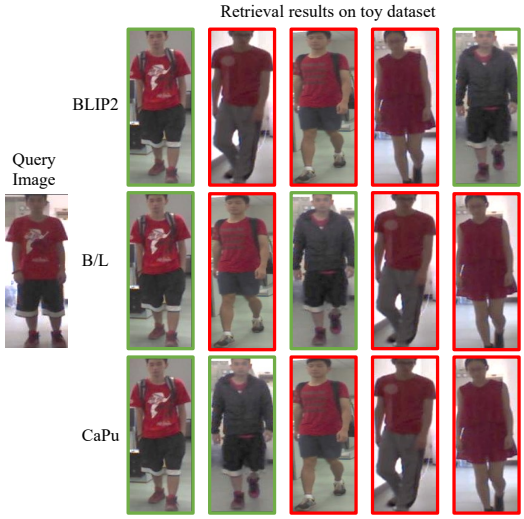


Figure 11 Visualization of the retrieval result on the toy datasets with only 5 images. The top row shows the results produced by BLIP2 (Li et al, 2023a); The second row shows the baseline results; the bottom row shows the results of CaPu. Green boxes denote the correct retrievals, and red boxes represent the incorrect ones.

results reveals noteworthy differences. As shown in the first row, the baseline method tends to retrieve images with superficial resemblances to the query image. This suggests a bias towards non-essential features like clothing, which can negatively impact retrieval accuracy, particularly in challenging scenarios such as occlusions or back views, where facial and upper-body information is limited.

In more detail, under the frontal view (Figure10 (a)), the baseline retrieves images with similar clothing patterns but mismatched identities, such as retrieving individuals wearing similar shirts, while CaPu correctly identifies matches by focusing on intrinsic identity features like body shape and structure. Under occlusion (Figure 10 (b)), the baseline model fails to retrieve accurate matches when key regions are hidden, instead relying on partial cues such as visible footwear. In contrast, CaPu captures discriminative identity details from less-occluded body parts to retrieve correct matches. In side-view and back-view scenarios (Figures.10 (c) and 10 (d)), where facial information is unavailable, the baseline often retrieves results based on non-discriminative features like pants or shoes. In contrast, CaPu accurately captures identity-relevant features, even in these challenging conditions.

Furthermore, in Figure 11 the toy dataset includes the top 5 gallery images with the highest similarity scores to the query image by BLIP2 (Li et al, 2023a), as illustrated in the first row. These images provide a direct and intuitive comparison among BLIP2, baseline methods, and CaPu, thereby elucidating their relative performance. The second row depicts the retrieval results obtained by the baseline method. Notably, the baseline method captures certain identity cues, evident in the improvement in the R@3 image compared to BLIP2. However, the baseline model remains susceptible to clothing interference. Similar clothing colors, identical pants, and shoes contribute to potential confusion in the retrieval results. In contrast, the third row showcases CaPu's retrieval results, demonstrating the impact of feature purification. CaPu, with its refined features, yields more precise retrieval results. The influence of clothing interference is notably reduced, resulting in a more accurate representation of intrinsic identity cues. This analysis of retrieval results underscores CaPu's efficacy in mitigating the impact of clothing-related spurious associations and highlights the potential for improved performance in real-world scenarios.

These results demonstrate CaPu's ability to purify identity features and mitigate spurious associations introduced by clothing or other irrelevant factors. Across all scenarios, CaPu consistently outperforms the baseline, achieving more robust retrievals by leveraging body shape and subtle identity cues.

5 Conclusion

This work presents a Causality-based Purification (CaPu) model for Cloth-Changing Person Re-Identification (CC-ReID). Theoretically, CaPu incorporates the generalization ability of Vision-Language Pretraining (VLP) models as clothing indicators to provide clothing semantics indications as the effective acquisition of prior knowledge. Furthermore, CaPu redefines the challenge of breaking spurious associations between identity and clothing information as a causality purification problem, which effectively utilizes the obtained prior knowledge. Specifically, the clothing annotations are accurately captured by the VLP indication pipeline, and the identity information from the expert model is purified by diminishing the impact of clothing on the learned visual representation from two causal perspectives: Consistency Treatment Effects (CTE) and Distinctiveness Treatment Effects (DTE). Comprehensive experiments across three CC-ReID datasets, PRCC, LTCC, and VC-Cloth, demonstrate that CaPu outperforms state-of-the-art methods.

Limitation Using large-scale VLP models requires more computational resources than expert models, which could be a limitation in certain settings. To mitigate this, we've structured the VLP processing in a modular way, enabling batch processing to reduce overhead. However, the computational demands remain, especially in resource-constrained environments. As large models continue to be streamlined, more efficient solutions may emerge.

References

Bai Y, Cao M, Gao D, et al (2023) Rasa: Relation and sensitivity aware representation learning for text-based person search. In: Proceeding of the International Joint Conferences on Artificial Intelligence, pp 555–563

Bannur S, Hyland SL, Liu Q, et al (2023) Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 15016–15027

Bansal V, Foresti GL, Martinel N (2022) Cloth-changing person re-identification with self-attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, pp 602–610

Barbosa IB, Cristani M, Bue AD, et al (2012) Re-identification with RGB-D sensors. In: European conference on computer vision Workshops, Springer, pp 433–442

Chen C, Ye M, Qi M, et al (2022a) Structure-aware positional transformer for visible-infrared person re-identification. IEEE Transactions on Image Processing 31:2352–2364

Chen C, Ye M, Jiang D (2023a) Towards modality-agnostic person re-identification with descriptive query. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 15128–15137

Chen J, Jiang X, Wang F, et al (2021) Learning 3d shape feature for texture-insensitive person re-identification. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 8146–8155

Chen J, Gao Z, Wu X, et al (2023b) Meta-causal learning for single domain generalization. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 7683–7692

Chen T, Ding S, Xie J, et al (2019) Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 8351–8361

Chen Z, Li G, Wan X (2022b) Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In: Proceedings of the ACM International Conference on Multimedia, pp 5152–5161

Ci Y, Wang Y, Chen M, et al (2023) Unihcp: A unified model for human-centric perceptions. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 17840–17852

Cui Z, Zhou J, Peng Y, et al (2023) Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 33(8):4415–4428

Dash S, Balasubramanian VN, Sharma A (2022) Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In: Proc. IEEE/CVF Winter Conference on Appl. Comput. Vis., pp 3879–3888

Dou ZY, Kamath A, Gan Z, et al (2022) Coarse-to-fine vision-language pre-training with fusion in the backbone. Advances in Neural Information Processing Systems 35:32942–32956

Du Y, Wei F, Zhang Z, et al (2022) Learning to prompt for open-vocabulary object detection with vision-language model. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 14084–14093

Fenton NE, Neil M, Constantinou AC (2020) The book of why: The new science of cause and effect, judea pearl, dana mackenzie. basic books (2018). Artif Intell 284:103286

Fu Y, Wei Y, Zhou Y, et al (2019) Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 8295–8302

Gao Z, Wei H, Guan W, et al (2022) Multigranular visual-semantic embedding for cloth-changing person re-identification. In: Proceedings of the ACM International Conference on Multimedia, pp 3703–3711

Gu X, Chang H, Ma B, et al (2022) Clothes-changing person re-identification with RGB modality only. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 1050–1059

Guo R, Cheng L, Li J, et al (2020) A survey of learning causality with data: Problems and methods. ACM Comput Surv 53(4):75:1–75:37

Han K, Gong S, Huang Y, et al (2023) Clothing-change feature augmentation for person re-identification. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 22066–22075

He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 770–778

He S, Luo H, Wang P, et al (2021) Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 15013–15022

He S, Chen W, Wang K, et al (2023) Region generation and assessment network for occluded person re-identification. IEEE Transactions on Information Forensics and Security 19:120–132

Hong P, Wu T, Wu A, et al (2021) Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 10513–10522

Hou R, Ma B, Chang H, et al (2019) Interaction-and-aggregation network for person re-identification. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 9317–9326

Huang H, Li D, Zhang Z, et al (2018) Adversarially occluded samples for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5098–5107

Huang Y, Wu Q, Xu J, et al (2021) Clothing status awareness for long-term person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 11875–11884

Jia X, Zhong X, Ye M, et al (2022) Complementary data augmentation for cloth-changing person re-identification. IEEE Transactions on Image Processing 31:4227–4239

Jin X, He T, Zheng K, et al (2022) Cloth-changing person re-identification from A single image with gait prediction and regularization. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 14258–14267

Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations, pp 1–15

Kirillov A, Mintun E, Ravi N, et al (2023) Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 4015–4026

Kuang K, Li L, Geng Z, et al (2020) Causal inference. Eng 6(3):253–263

Kweon HJ, Cho D (2023) Cloth-changing person re-identification with noisy patch filtering. IEEE Signal Processing Letters 30:334–338

Li J, Li D, Xiong C, et al (2022a) BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, PMLR, pp 12888–12900

Li J, Li D, Savarese S, et al (2023a) BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, PMLR, pp 19730–19742

Li J, Niu L, Zhang L (2023b) Knowledge proxy intervention for deconfounded video question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 2782–2793

Li P, Xu Y, Wei Y, et al (2020) Self-correction for human parsing. IEEE Transactions on Pattern Analysis & Machine Intelligence 44(6):3260–3271

Li S, Chen H, Yu S, et al (2023c) Cocas+: Large-scale clothes-changing person re-identification with clothes templates. IEEE Transactions on Circuits and Systems for Video Technology 33(4):1839–1853

Li S, Sun L, Li Q (2023d) Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 1405–1413

Li W, Hou S, Zhang C, et al (2023e) An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 13824–13833

Li X, Liu B, Lu Y, et al (2022b) Cloth-aware center cluster loss for cloth-changing person re-identification. In: Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision, Springer, pp 527–539

Li X, Lu Y, Liu B, et al (2022c) Counterfactual intervention feature transfer for visible-infrared person re-identification. In: European conference on computer vision, Springer, pp 381–398

Li X, Lu Y, Liu B, et al (2023f) Clothes-invariant feature learning by causal intervention for clothes-changing person re-identification. CoRR abs/2305.06145

Li YJ, Weng X, Kitani KM (2021) Learning shape representations for person re-identification under clothing change. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2432–2441

Liu F, Kim M, Gu Z, et al (2023a) Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 19617–19626

Liu F, Ye M, Du B (2023b) Dual level adaptive weighting for cloth-changing person re-identification. IEEE Transactions on Image Processing 32:5075–5086

Liu J, Shen Z, Cui P, et al (2021) Stable adversarial learning under distributional shifts. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 8662–8670

Liu R, Liu H, Li G, et al (2022) Contextual debiasing for visual recognition with causal mechanisms. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 12745–12755

Liu S, Zeng Z, Ren T, et al (2023c) Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. CoRR abs/2303.05499

Liu Y, Ge H, Wang Z, et al (2023d) Clothes-changing person re-identification via universal framework with association and forgetting learning. IEEE Transactions on Multimedia pp 1–14

Lu A, Zhang Z, Huang Y, et al (2024) Illumination distillation framework for nighttime person re-identification and a new benchmark. IEEE Transactions on Multimedia 26:406–419

Lv F, Liang J, Li S, et al (2022) Causality inspired representation learning for domain generalization. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 8046–8056

Mahajan D, Tople S, Sharma A (2021) Domain generalization using causal matching. In: International Conference on Machine Learning, PMLR, pp 7313–7324

Mao C, Xia K, Wang J, et al (2022) Causal transportability for visual recognition. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 7521–7531

Miao J, Chen C, Liu F, et al (2023) Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 21426–21437

Ming Y, Cai Z, Gu J, et al (2022) Delving into out-of-distribution detection with vision-language representations. Advances in Neural Information Processing Systems 35:35087–35102

Nguyen VD, Khaldi K, Nguyen D, et al (2024) Contrastive viewpoint-aware shape learning for long-term person re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 1041–1049

Niu Y, Tang K, Zhang H, et al (2021) Counterfactual VQA: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 12700–12710

Ouyang C, Chen C, Li S, et al (2022) Causality-inspired single-source domain generalization for medical image segmentation. IEEE Transactions on Medical Imaging 42(4):1095–1106

Pearl J (2010) Causal inference. In: JMLR Causality: Objectives and Assessment, pp 39–58

Pearl J (2013) Direct and indirect effects. arXiv:13012300

Peng C, Wang B, Liu D, et al (2024) Masked attribute description embedding for cloth-changing person re-identification. IEEE Transactions on Multimedia

Qian X, Wang W, Zhang L, et al (2020) Long-term cloth-changing person re-identification. In: Proceedings of the Asian Conference on Computer Vision, pp 71–88

Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, PMLR, vol 139. PMLR, pp 8748–8763

Rao Y, Chen G, Lu J, et al (2021) Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1005–1014

Ren T, Liu S, Zeng A, et al (2024) Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:240114159

Selvaraju RR, Cogswell M, Das A, et al (2020) Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision 128(2):336–359

Shao Z, Zhang X, Ding C, et al (2023) Unified pre-training with pseudo texts for text-to-image person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 11174–11184

Shi W, Liu H, Liu M (2022) IRANet: identity-relevance aware representation for cloth-changing person re-identification. Image Vis Comput 117:104335

Shu X, Li G, Wang X, et al (2021a) Semantic-guided pixel sampling for cloth-changing person re-identification. IEEE Signal Processing Letters 28:1365–1369

Shu X, Wang X, Zang X, et al (2021b) Large-scale spatio-temporal person re-identification: Algorithms and benchmark. IEEE Transactions on Circuits and Systems for Video Technology 32(7):4390–4403

Siddiqui N, Croitoru FA, Nayak GK, et al (2024) Dlcr: A generative data expansion framework via diffusion for clothes-changing person re-id. arXiv preprint arXiv:241107205

Singh KK, Lee YJ (2017) Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, pp 3544–3553

Somers V, De Vleeschouwer C, Alahi A (2023) Body part-based representation learning for occluded person re-identification. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 1613–1623

Song C, Huang Y, Ouyang W, et al (2018) Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1179–1188

Song S, Wan J, Yang Z, et al (2022) Vision-language pre-training for boosting scene text detectors. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 15681–15691

Sun S, Zhi S, Liao Q, et al (2023) Unbiased scene graph generation via two-stage causal modeling. IEEE Transactions on Pattern Analysis & Machine Intelligence 45(10):12562–12580

Sun Y, Zheng L, Yang Y, et al (2018) Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In: European conference on computer vision, Springer, pp 501–518

Tang K, Niu Y, Huang J, et al (2020) Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 3713–3722

Tang S, Chen C, Xie Q, et al (2023) Humanbench: Towards general human-centric perception with projector assisted pretraining. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 21970–21982

Wan F, Wu Y, Qian X, et al (2020) When person re-identification meets changing clothes. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference Workshop, pp 830–831

Wang P, Bai S, Tan S, et al (2024a) Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. CoRR abs/2409.12191

Wang Q, Qian X, Li B, et al (2024b) Exploring fine-grained representation and recomposition for cloth-changing person re-identification. IEEE Transactions on Information Forensics and Security

Wu J, Yang Y, Lei Z, et al (2023a) Camera-aware representation learning for person re-identification. Neurocomputing 518:155–164

Wu L, Liu D, Zhang W, et al (2022) Pseudo-pair based self-similarity learning for unsupervised person re-identification. IEEE Transactions on Image Processing 31:4803–4816

Wu Y, Wei P, Lin L (2023b) Scene graph to image synthesis via knowledge consensus. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 2856–2865

Xie D, Liu L, Zhang S, et al (2023) A unified multi-modal structure for retrieving tracked vehicles through natural language descriptions. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 5418–5426

Xue D, Qian S, Xu C (2023) Variational causal inference network for explanatory visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 2515–2525

Yan B, Pei M (2022) Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 2982–2990

Yan S, Dong N, Zhang L, et al (2023) Clip-driven fine-grained text-image person re-identification. IEEE Transactions on Image Processing 32:6032–6046

Yan Y, Yu H, Li S, et al (2022) Weakening the influence of clothing: Universal clothing attribute disentanglement for person re-identification. In: Proceeding of the International Joint Conferences on Artificial Intelligence, pp 1523–1529

Yang K, Tian X (2023) Domain-class correlation decomposition for generalizable person re-identification. IEEE Transactions on Multimedia 25:3386–3396

Yang M, Liu F, Chen Z, et al (2021a) CausalVAE: disentangled representation learning via neural structural causal models. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 9593–9602

Yang Q, Wu A, Zheng W (2021b) Person re-identification by contour sketch under moderate clothing change. IEEE Transactions on Pattern Analysis & Machine Intelligence 43(6):2029–2046

Yang S, Kang B, Lee Y (2022a) Sampling agnostic feature representation for long-term person re-identification. IEEE Transactions on Image Processing 31:6412–6423

Yang Z, Zhong X, Liu H, et al (2022b) Attentive decoupling network for cloth-changing re-identification. In: Proceedings of the IEEE International conference on Multimedia and Expo, pp 1–6

Yang Z, Lin M, Zhong X, et al (2023a) Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 1472–1481

Yang Z, Zhong X, Zhong Z, et al (2023b) Win-win by competition: Auxiliary-free cloth-changing person re-identification. IEEE Transactions on Image Processing 32:2985–2999

Yao Y, Yu T, Zhang A, et al (2024) Minicpm-v: A GPT-4V level MLLM on your phone. CoRR abs/2408.01800

Ye M, Shen J, Lin G, et al (2021) Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis & Machine Intelligence 44(6):2872–2893

Yu H, Liu B, Lu Y, et al (2022a) Multi-view geometry distillation for cloth-changing person reid. In: Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision, Springer, pp 29–41

Yu S, Li S, Chen D, et al (2020) COCAS: A large-scale clothes changing person dataset for re-identification. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 3397–3406

Yu Y, Zhan F, Wu R, et al (2022b) Towards counterfactual image manipulation via clip. In: Proceedings of the ACM International Conference on Multimedia, pp 3637–3645

Zang C, Wang H, Pei M, et al (2023) Discovering the real association: Multimodal causal reasoning in video question answering. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 19027–19036

Zhang G, Luo Z, Chen Y, et al (2022) Illumination unification for person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 32(10):6766–6777

Zhang G, Liu J, Chen Y, et al (2023a) Multi-biometric unified network for cloth-changing person re-identification. IEEE Transactions on Image Processing 32:4555–4566

Zhang K, Yang Y, Yu J, et al (2024) Multi-task paired masking with alignment modeling for medical vision-language pre-training. IEEE Transactions on Multimedia 26:4706–4721

Zhang X, Cui P, Xu R, et al (2021) Deep stable learning for out-of-distribution generalization. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 5372–5382

Zhang YF, Zhang Z, Li D, et al (2023b) Learning domain invariant representations for generalizable person re-identification. IEEE Transactions on Image Processing 32:509–523

Zheng Z, Yang X, Yu Z, et al (2019) Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference, pp 2138–2147

Zhong Z, Zheng L, Kang G, et al (2020) Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 13001–13008

Zhou K, Yang Y, Cavallaro A, et al (2019) Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 3701–3711

Zhou Z, Liu H, Shi W, et al (2022) A cloth-irrelevant harmonious attention network for cloth-changing person re-identification. In: Proc. IEEE Int. Conf. Pattern Recog., pp 989–995

Zhu D, Chen J, Shen X, et al (2024) MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: Proceedings of the International Conference on Learning Representations, pp 1–17

Zhu K, Guo H, Liu Z, et al (2020) Identity-guided human semantic parsing for person re-identification. In: European conference on computer vision, Springer, pp 346–363