Article

Not peer-reviewed version

Virtual Try-On–Based Data Augmentation for Robust Person Re-Identification in Emergency Surveillance Scenarios

Pei Wang , Jiaming Liu , Yuyao Cao , Hui Zhang *

Posted Date: 26 January 2026

doi: 10.20944/preprints202601.1873.v1

Keywords: person re-identification; emergency surveillance; data augmentation; virtual try-on

 doi:10.20944/preprints202601.1873.v1

Article

Virtual Try-On–Based Data Augmentation for Robust Person Re-Identification in Emergency Surveillance Scenarios

Pei Wang [1], Jiaming Liu [1], Yuyao Cao [1] and Hui Zhang [1,*]

School of Safety Science, Tsinghua University, Beijing, China
* Correspondence: zhhui@tsinghua.edu.cn

Abstract

Person Re-identification (person Re-ID) plays an important role in dynamic evacuation path planning and person tracking in emergency scenarios. However, its robustness is severely challenged in such conditions, where persons' appearances may change rapidly due to stress responses or environmental interventions. Meanwhile, privacy regulations and data access constraints limit the availability of long-term surveillance data, hindering the generalization capability of re-identification models. Virtual try-on technologies offer a promising means of enriching appearance diversity under limited data conditions. In this study, a virtual try-on–based data augmentation method for person Re-ID is developed. To address inaccurate clothing mask extraction caused by low image resolution, occlusions, and complex backgrounds, the original mask generation module used in existing virtual try-on pipelines is replaced with a composite framework integrating Grounding DINO and the Segment Anything Model (SAM). The proposed framework enables precise extraction of clothing regions using text-based prompts, through which appearance-diverse person images are generated. Extensive comparative experiments and multi-level analyses demonstrate that the generated images exhibit high visual realism, preserve identity-related information, and do not introduce systematic distribution shifts. Controlled experiments on a ResNet-50–based benchmark further confirm that the proposed data augmentation strategy consistently improves re-identification performance.

Keywords: person re-identification; emergency surveillance; data augmentation; virtual try-on

1. Introduction

The integration of person re-identification and person evacuation research has recently emerged as an interdisciplinary direction. Person re-identification (Re-ID) [1–3] focuses on matching the same individual across non-overlapping camera views. Under emergency conditions, person Re-ID systems can be used in real-time tracking of specific individuals, such as older adults or children. Based on location information, timely guidance toward safe areas can be provided. Meanwhile, crowd density in different zones can be estimated using multi-camera video streams. Such information is essential for adaptive evacuation guidance. Under normal operating conditions, person trajectories can be reconstructed over time. Congestion-prone areas within buildings can therefore be identified in advance. Potential safety hazards may be mitigated accordingly. In addition, mobility differences among demographic groups can be analyzed, which provides valuable support for safety assessment and evacuation planning.

Despite its potential in evacuation scenarios, the robustness of person Re-ID is strongly affected under emergency conditions. Rapid environmental changes may lead to substantial appearance variations within short time intervals. Clothing changes and the use of protective coverings further increase this variability, by which the performance of existing Re-ID models is often degraded. Moreover, the training of person Re-ID models is highly dependent on large-scale surveillance data collected in real

 doi:10.20944/preprints202601.1873.v1

environments. In practice, due to privacy regulations and ethical constraints, obtaining long-term, cross-scenario datasets is strictly limited. Therefore, it is difficult to obtain images of the same person with different apperances, which becomes a major bottleneck in improving the robustness and generalization ability of models.

Clothing-change person re-identification (CC-Person Re-ID) [12–14] is a subtask of person Re-ID that addresses appearance variations caused by changes in clothing. It focuses on improving model performance when persons undergo significant changes in appearance or attire. Existing studies primarily investigate this problem from two perspectives: model-level improvements and data augmentation strategies.

From a data augmentation perspective, existing approaches mainly follow two directions. The first focuses on constructing datasets that contain multiple clothing appearances for the same identity, such as DeepChange [13], LTCC [12], and PRCC [14]. Although these datasets improve model performance to some extent, their construction is labor-intensive and the resulting appearance diversity remains limited. The second direction leverages generative techniques, such as modifying clothing color or altering person pose, to increase data diversity. However, due to limitations inherent in the generation techniques at the time, the synthesized images often lack sufficient realism and diversity.

In recent years, generative technologies, such as generative adversarial networks (GANs) [15] and diffusion models [16], have made substantial progress in producing content with high diversity and realism. One representative application is virtual try-on [17–20] in e-commerce, where generative models enable customers to virtually "wear" selected garments using uploaded personal images, as illustrated in Figure 1.

Figure 1. Illustration of Virtual Try-On.

Inspired by advances in virtual try-on systems and motivated by the need to alleviate data scarcity in person Re-ID, the application of virtual try-on techniques to data augmentation is explored in this study. A virtual try-on–based augmentation approach is developed for person Re-ID. By generating clothing-change images with high visual realism and rich appearance diversity, this approach aims to enhance model robustness under significant person appearance changes.

However, virtual try-on in e-commerce is typically developed under highly controlled imaging conditions. Model and garment images are captured with high resolution and uniform lighting. Human poses are simple, and viewing angles are near horizontal. In contrast, person images for person Re-ID are acquired by surveillance cameras. They often suffer from motion blur, low resolution, and limited image clarity. Body poses are complex. Scenes are unconstrained, and occlusions from surrounding objects and carried items are common. Moreover, cameras are usually installed at elevated positions, producing top-down or oblique views, as shown in Figure 2 . These factors together pose substantial challenges for clothing-change generation in surveillance imagery.

Figure 2. Common Issues in Person Images for person Re-ID.

 doi:10.20944/preprints202601.1873.v1

When virtual try-on techniques are applied to person images captured in surveillance scenarios, the accuracy of clothing region masks plays a critical role in determining the final synthesis quality. If the original mask generation module in IDM-VTON [17] (a virtual try-on model used in this article) are directly adopted, the resulting masks are often inaccurate due to the inherent characteristics of surveillance imagery. As a result, clothing replacement frequently fails, and the generated images are unsuitable for data augmentation.

To facilitate the use of virtual try-on for data augmentation in person Re-ID, a prompt-based automatic clothing mask generation (PACMG) method is proposed and integrated into the virtual try-on pipeline. The module uses prompt-driven segmentation to extract clothing regions of a specified target person in both single-person and multi-person scenes. Different garment descriptions can be assigned as prompts. Experiments show that PACMG produces higher-quality masks and achieves more realistic clothing replacement.

Beyond visual assessment, a downstream evaluation is conducted to examine whether the augmented data lead to improvements in person Re-ID performance. A ResNet-50–based model is used as the baseline, and the proposed replacement method is applied to augment the training set. A tiered augmentation strategy is adopted to combine original and augmented samples during training. Results indicate consistent gains across multiple retrieval metrics, confirming the effectiveness of the proposed clothing-based augmentation.

2. Related Works

2.1. Person Re-ID in Surveillance and Emergency Scenarios

In public safety and emergency scenarios, the primary concern is person. Person Re-ID (Re-ID) [2] associates the same individual across multiple cameras deployed in different locations, making continuous localization and tracking of person possible. This capability makes Re-ID naturally suitable for security monitoring and evacuation management. In security systems, Re-ID has been widely studied for rapid target localization and cross-camera tracking. Most studies take the task as identity classification and metric learning, with performance improved through stronger network backbones [9–11] and optimization strategies [1]. In this study, ResNet-50 is adopted as the baseline framework.

In practical surveillance scenarios, Re-ID performance is often degraded by lighting variation, occlusion, long time tracking, and appearance changes. Several studies have therefore focused on infrared-based identification, long-term tracking [12], and clothing-change Re-ID [12–14]. However, these approaches, while improving accuracy in their targeted problems, are often only effective for isolated or limited scenarios.

Given its relevance to evacuation tasks, Re-ID has also been explored in emergency management. For example, multi-camera systems combined with Re-ID have been used for indoor localization, trajectory tracking, and crowd counting, with the resulting data mapped to building digital twins to support abnormal behavior analysis and evacuation planning [6]. Human behavior detection dataset (HBDset) [5] have been constructed to improve the detection and tracking of vulnerable groups, providing important data support for person-centric safety applications. In addition, a study has further integrated Re-ID with sensors and edge computing platforms to enable real-time crowd monitoring, fire detection, and evacuation route generation [7].

Overall, although direct applications of Re-ID to evacuation remain limited, its potential for continuous localization, trajectory reconstruction, and key-person identification is significant.

2.2. Data Augmentation for Person Re-ID

Data augmentation is an effective technique for alleviating data imbalance and data scarcity by increasing sample diversity, thereby improving model generalization performance. In the field of person Re-ID, commonly used basic augmentation strategies include random flipping, rotation, cropping, scaling, color jittering, random erasing, and occlusion. These operations fully exploit existing data and help models learn more robust and discriminative features. Beyond basic augmentation techniques, another

 doi:10.20944/preprints202601.1873.v1

line of research focuses on generating new training samples. For example, ClothMix [21] enhances clothing texture diversity by exchanging the statistical moments of clothing-related feature regions across different samples in the feature space. Person features are divided into identity-related regions (e.g., the head) and clothing-related regions (e.g., garments). Clothing features from different samples within a batch are then randomly mixed, producing diversified clothing styles while preserving identity consistency.

With the development of generative models, several studies have leveraged generative adversarial networks (GANs) [22] to synthesize new person images. These methods typically perform pose transformation, appearance swapping, clothing color transfer [23–26], or style editing. In some cases, background replacement is also applied [27]. Such approaches can generate appearance-diverse images while maintaining identity information. However, notable limitations remain. Some methods effectively replicate existing appearances by jointly transferring both upper- and lower-body attributes, which restricts localized appearance editing and limits diversity. In addition, the realism of the generated images is often constrained by the performance of the generative models.

Driven by the growth of e-commerce platforms and increasing demand for personalized services, virtual try-on has recently become an active research topic [30]. In particular, diffusion-based models have enabled the development of high-quality virtual try-on systems. These models generate high-resolution images of individuals wearing different garments based on personal photos or selected templates, offering strong flexibility and photo-realistic results. Inspired by these advances, this work explores the use of diffusion-based virtual try-on techniques for data augmentation in person Re-ID.

2.3. Image-Based Virtual Try-On

Image-based virtual try-on [36] refers to a class of generative methods that synthesize a target garment onto a person or model image to simulate real-world try-on effects [17,18,28]. This technology allows users to visually preview how clothing would appear when worn and is widely used to support online purchasing decisions. Early virtual try-on systems were mainly based on generative adversarial networks (GANs) [33,34]. However, GAN-based methods often suffer from limited visual realism, geometric distortions, and weak generalization, especially under diverse poses and complex scenes.

Another typical solution is the warping-based pipeline [31,32]. In this framework, the target garment is first geometrically deformed and aligned according to the posture of the target person. A generative network is then applied to synthesize the final image. Warping-based methods usually require two separate stages to complete the try-on process.

Recent virtual try-on approaches are primarily built upon diffusion models, which have significantly improved visual realism and generalization ability. From the perspective of garment source, existing methods can be divided into two categories. Garment-to-person try-on [17,18,20] synthesizes a standalone clothing image onto a given person image and requires paired garment and person inputs. In contrast, person-to-person try-on [19,35] transfers garments worn by one individual to another person and therefore requires two person images with different clothing styles.

From the viewpoint of application scenarios, virtual try-on methods can be classified into in-shop try-on [17,18] and try-on in the wild [28,29,37]. In-shop try-on usually assumes simple and uniform backgrounds, such as white or solid-color backdrops, and is widely adopted in e-commerce. By contrast, try-on in the wild targets real-world scenes with complex and variable backgrounds, making clothing replacement considerably more challenging.

Depending on whether explicit clothing masks are used, virtual try-on methods can be further categorized into mask-based [17,18] and mask-free [29,38–41] approaches. In mask-based methods, the accuracy of clothing region masks directly determines the quality of the synthesized results. To overcome this limitation, recent studies have explored mask-free frameworks, some of which achieve performance comparable to or even better than mask-based approaches. It should be noted, however, that several mask-free models still rely on mask-based techniques during training to generate paired supervision data. Therefore, mask-based virtual try-on remains an important research direction. Based on the above analysis, this work adopts a mask-based virtual try-on framework to study person clothing replacement in surveillance scenarios.

 doi:10.20944/preprints202601.1873.v1

3. Methodology

Virtual try-on serves different purposes in e-commerce and person Re-ID. In e-commerce, the primary goal is to provide visually realistic and personalized clothing previews. Therefore, preserving the fine details of the garments and achieving high visual fidelity is crucial. In person Re-ID, clothing replacement is mainly used for data augmentation. The objective is to increase sample size, balance identity distributions, and enhance appearance diversity to improve model generalization. In this context, preserving the fine texture of clothing is less important. Instead, consistency of identity after clothing replacement and overall visual realism are even more crucial. In short, e-commerce applications emphasize clothing details and photorealistic appearance, while person Re-ID prioritizes identity preservation, appearance diversity, and realism.

In addition, the difficulties encountered in the two scenarios are different significantly: images in e-commerce applications typically have high resolution, clear backgrounds, and the person in the images have simple poses, without complex perspectives or accessories, and the clothing area is unobstructed. In the field of person Re-ID, images often suffer from low resolution, motion blur, varied poses, top-down perspectives, complex backgrounds, and partial occlusions. These factors pose a significant challenge to pedestrians changing clothes in surveillance scenarios.

Recent advances in diffusion-based image generation have greatly improved the realism of virtual try-on. Several effective models have been proposed, among which IDM-VTON shows strong performance. In this work, IDM-VTON is adopted as the base clothing replacement model. Task-oriented modifications are further introduced to adapt it to surveillance-based person Re-ID.

This section presents the overall methodology of the proposed data augmentation framework for person Re-ID. As illustrated in Figure 3, the framework consists of three main stages and one application. First, a prompt-based automatic clothing mask generation (PACMG) method is introduced to extract accurate clothing regions from surveillance images. Second, the generated masks are integrated into the IDM-VTON model to perform clothing replacement and synthesize appearance-diverse person images. Third, the augmented data are combined with the original dataset according to a tiered data augmentation strategy. Then the augmented images are used to train a person Re-ID model. The trained model is then applied to downstream re-identification tasks. Each stage of the framework is described in detail in the following subsections.

Figure 3. The framework of the method proposed in this paper.

 doi:10.20944/preprints202601.1873.v1

3.1. IDM-VTON

IDM-VTON [17] is a diffusion-based virtual try-on method developed by the Korea Advanced Institute of Science and Technology (KAIST). It integrates an IP-Adapter and adopts a dual-branch architecture composed of TryonNet and GarmentNet. TryonNet processes the person image, while GarmentNet extracts low-level garment details. The IP-Adapter captures high-level semantic information of the target clothing and facilitates effective feature fusion during synthesis.

IDM-VTON shows strong performance in both in-shop and in-the-wild scenarios. It achieves an FID score of 8.64 on the DressCode dataset and 6.29 on the VITON-HD dataset, indicating high visual quality and realism. Motivated by this performance, IDM-VTON is adopted in this work as the base model for clothing replacement on person images in person Re-ID datasets for data augmentation.

The clothing replacement process of IDM-VTON is illustrated in Figure 4. As shown in the figure, multiple inputs are required before clothing replacement, including garment images, person images, human pose information (generated by Openpose [47]) , human parsing results (generated by SCHP [46]), 3D body surface representations (generated by Densepose [48]), region masks, and textual garment descriptions. Among these inputs, the upper-body clothing mask is generated from the person image using a human parsing map and pose keypoints. Clothing regions are first extracted based on semantic labels and then refined with pose information around the shoulders and arms to improve boundary accuracy, resulting in a binary mask for subsequent virtual try-on.

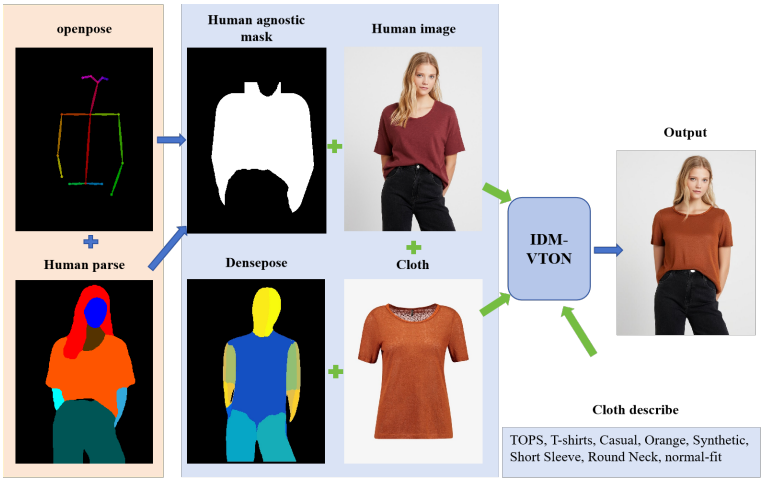
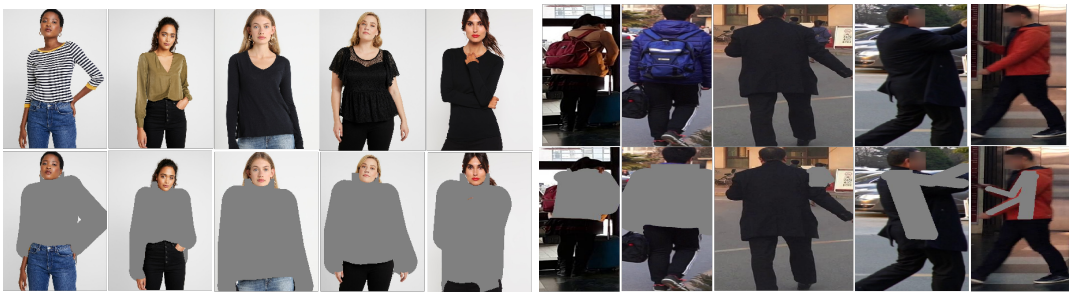


Figure 4. The clothing replacement process of IDM-VTON.

3.2. Prompt-Based Automatic Clothing Mask Generation for Person Re-ID

The above mask generation approach performs well when person images are clear, poses are simple, and clothing regions are minimally occluded, as shown in Figure 5(**a**). However, in real-world surveillance scenes, images often contain complex poses, occlusions from carried items (e.g., backpacks), background clutter, and interference from nearby persons. Under these conditions, mask generation based on human parsing and pose keypoints is prone to errors. Non-clothing regions may be mistakenly included in the mask, leading to loss of appearance details and reduced visual realism in subsequent clothing replacement, as illustrated in Figure 5(**b**).

 doi:10.20944/preprints202601.1873.v1

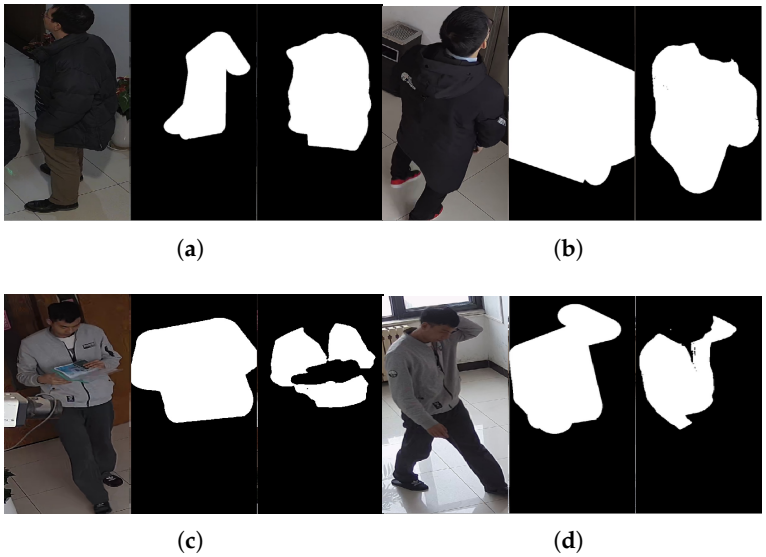


(**a**) (**b**)

Figure 5. Application of Parsing-Based Mask Generation in (**a**) Virtual Try-On and (**b**) Person Re-ID. (These images come from ICFG-PEDES [8])

To overcome these limitations, an open-vocabulary segmentation framework combining Grounding DINO [42] and the Segment Anything Model (SAM) [45] is adopted to directly segment clothing regions using text prompts. Grounding DINO localizes target objects based on textual descriptions and outputs bounding boxes, while SAM generates accurate segmentation masks from such prompts. By integrating the two models, clothing regions can be automatically segmented using garment category prompts (e.g., coat or jacket), enabling batch-wise mask generation.

Using text-based garment prompts allows the proposed method to better align with the actual visual appearance of clothing and suppress interference from occluding objects, resulting in higher-quality masks. Figure 6 compares the two approaches. Each group shows the original image, the mask generated by human parsing and pose keypoints, and the mask generated using the textual prompt "coat".



(**a**) (**b**)

(**c**) (**d**)

Figure 6. Comparison of Clothing Masks Generated by Pose–Parsing and Grounding DINO + SAM Methods Under Challenging Conditions (**a**) Side View. (**b**) Top-Down View. (**c**) Occlusion. (**d**) Complex Poses.

Figure 6 shows that the masks generated by the composite model closely follow the shapes of person clothing and effectively separate clothing from non-clothing regions. This enables preservation of motion-related details and carried items in the original images, while avoiding unintended modification of non-clothing areas caused by inaccurate masks.

The proposed mask generation strategy also simplifies data preparation. Clothing masks are obtained using only garment category prompts such as coat or jacket. No additional pose estimation or human parsing is required, which reduces both preprocessing complexity and computational cost.

 doi:10.20944/preprints202601.1873.v1

A comparison of the required input data for clothing replacement before and after the proposed improvement is summarized in Table 1.

Table 1. Comparison of required inputs for clothing replacement before and after the proposed improvement.

Method	Open-Pose	Human Parsing	Human-Agnostic Mask	Dense-Pose	Garment Description	Location Prompt	Person Image	Garment Image
IDM-VTON	✓	✓	✓	✓	✓	×	✓	✓
PACMG	×	×	✓	✓	✓	✓	✓	✓

The prompt-based automatic clothing mask generation (PACMG) method also offers strong flexibility and target specificity, which is particularly useful in multi-person scenes. The parsing- and pose-based methods may incorrectly assign masks to non-target individuals when multiple persons are present, just as the third group of Figure 5(**b**) shown, the background person is mistakenly selected.

In contrast, the prompt-driven approach allows target persons to be specified by modifying the textual prompts. As shown in the upper-left example of Figure 7, prompts such as "the black coat" and "the white coat" enable extraction of clothing masks for different individuals in the same image. This capability supports customized clothing replacement for multiple targets in crowded scenes.

Figure 7. Prompt-Based Automatic Clothing Mask Extraction in Multi-Person Scenes.

3.3. Tiered Data Augmentation Strategy

With the clothing masks obtained by the proposed method, IDM-VTON can be applied to perform clothing-based augmentation on person images captured in surveillance scenarios. The overall augmentation pipeline is illustrated in Figure 3. Since the ultimate goal of data augmentation is to improve model training, both the dataset size and its identity-level distribution have a direct impact on the learning performance. Based on statistical analysis of the original dataset, a tiered data augmentation strategy is designed to systematically expand the training data.

Assume that a person Re-ID dataset contains a total of N identities (IDs), and that the i-th identity has n_i images. The median number of images per identity is defined as m=$median$(n_i).

According to the sample size of each identity, all identities are divided into four groups:

Very few: $n_i < T_1$, Few: $T_1 \leq n_i < T_2$, Medium: $T_2 \leq n_i \leq m$, Dense: $n_i \geq m$,

where n_i denotes the number of images for the i-th identity and m is the median sample size of all identities in the training set. T_1 and T_2 are predefined thresholds used to categorize identities into

 doi:10.20944/preprints202601.1873.v1

different sample-density groups. Different augmentation targets are assigned to the four groups. The target number of images for each identity is defined as

$$T(n) = \begin{cases} \mathrm{m}, & n < 30, \\ 0.8 \times \mathrm{m}, & 30 \leq n < 60, \\ 0.9 \times \mathrm{m}, & 60 \leq n < \mathrm{m}, \\ 0, & n \geq \mathrm{m}, \end{cases} \tag{1}$$

where $T(n)$represents the target number of images for an identity after augmentation. Identities with fewer samples are augmented to higher proportions of the median, while no augmentation is applied when the sample size already exceeds the median.

4. Experiment and Analysis

This section presents the experimental evaluation of the proposed framework. First, comparative experiments on clothing replacement are conducted, and the results produced by the baseline and the improved methods are evaluated. Next, the validity of the augmented data is analyzed from two aspects: identity consistency with the original dataset and preservation of data distribution. Finally, the generated images are integrated into the original training set using the tiered augmentation strategy described in Section 3.3. Three groups of controlled experiments are then performed to verify whether the proposed clothing-based augmentation method improves person Re-ID performance.

4.1. Dataset and Evaluation Metrics

4.1.1. Virtual Try-On and Person Re-ID Datasets

The most widely used datasets for virtual try-on research include VITON-HD and DressCode. VITON-HD [44] contains 12,679 paired images of female models and upper-body garments at a resolution of 1024 × 768. All model images are captured in frontal poses. The dataset is split into 11,647 training pairs and 2,032 testing pairs.

DressCode [43] is larger in scale, with 53,792 model–garment pairs at the same resolution. It covers multiple garment categories, including upper-body clothing, dresses, and lower-body garments. Accordingly, DressCode is divided into three subsets based on garment type.

For person Re-ID experiments, a self-collected surveillance dataset is used. The dataset is constructed from 24-hour video recordings captured by five surveillance cameras in an indoor public area. It contains 159 identities and 43,742 person images with varying resolutions. The training set includes 131 identities with 39,150 images, while the test set contains 28 identities with 4,592 images. Due to privacy constraints, this dataset is restricted to internal use and cannot be publicly released.

A key feature of this dataset is the severe imbalance in the number of images per identity, ranging from a maximum of 2634 images to a minimum of 19. This long-tailed distribution makes the dataset highly suitable for evaluating data augmentation strategies in person re-identification.

In the clothing replacement experiments, the person images are taken from a self-collected surveillance dataset. The clothing images are from the VITON-HD test set, which contains 2032 women's garments. To achieve random pairing while maintaining reasonable visual consistency, only 1324 relatively neutral-style garments are retained in the experiment.

4.1.2. Evaluation Metrics

Clothing replacement for person Re-ID is an image-based generative task. The quality of generated images is evaluated with FID, LPIPS, SSIM, and CLIP Score.

- Fréchet Inception Distance (FID) measures the distributional discrepancy between generated images and real images. Lower values indicate closer alignment with the real data distribution;
- Learned Perceptual Image Patch Similarity (LPIPS) evaluates perceptual similarity in deep feature space, where lower scores correspond to higher visual similarity;

 doi:10.20944/preprints202601.1873.v1

- Structural Similarity Index Measure (SSIM) assesses structural similarity in terms of luminance, contrast, and texture, with values closer to 1 indicating higher similarity;
- CLIP Score measures semantic alignment between an image and its textual description, where higher values reflect better image–text consistency.

In this work, FID, LPIPS, SSIM, and a human-evaluated pass rate are jointly adopted to comprehensively assess the visual quality and practical usability of the generated clothing replacement results.

For person Re-ID, performance is evaluated using Rank-k accuracy and mean Average Precision (mAP).

Rank-k reports the percentage of queries for which at least one true match appears in the top-k retrieved results. mAP evaluates retrieval quality over the full ranked list by averaging the precision across all correct matches for each query, and then averaging over all queries.

Unless otherwise specified, Rank-1/5/10 and mAP are reported under the standard evaluation protocol. All evaluations are conducted in the cross-camera setting without re-ranking.

4.2. Comparison and Evaluation of Clothing Replacement Results

4.2.1. Clothing Replacement Results

To compare the proposed prompt-based automatic clothing mask generation (PACMG) method with the original IDM-VTON pipeline, two sets of clothing replacement data are constructed. The required inputs for both settings are summarized in Table 1, including person images, garment images, clothing masks, DensePose-based dense surface correspondences, and textual garment descriptions. The two settings differ only in the mask generation strategy: one uses human parsing and pose keypoints, while the other adopts the proposed prompt-based automatic clothing mask generation (PACMG) method.

IDM-VTON is applied to both datasets to perform clothing replacement. Qualitative results are shown in Figure 8. Under challenging surveillance conditions, images generated with the proposed method better preserve facial regions and fine-grained hand details. The overall visual appearance is also more natural. Although the replaced garments do not strictly reproduce the original textures, this behavior is acceptable for person Re-ID. In Re-ID tasks, preserving identity-related cues is more critical than maintaining garment detail. The proposed method therefore better satisfies the practical requirements of identity preservation.

 doi:10.20944/preprints202601.1873.v1

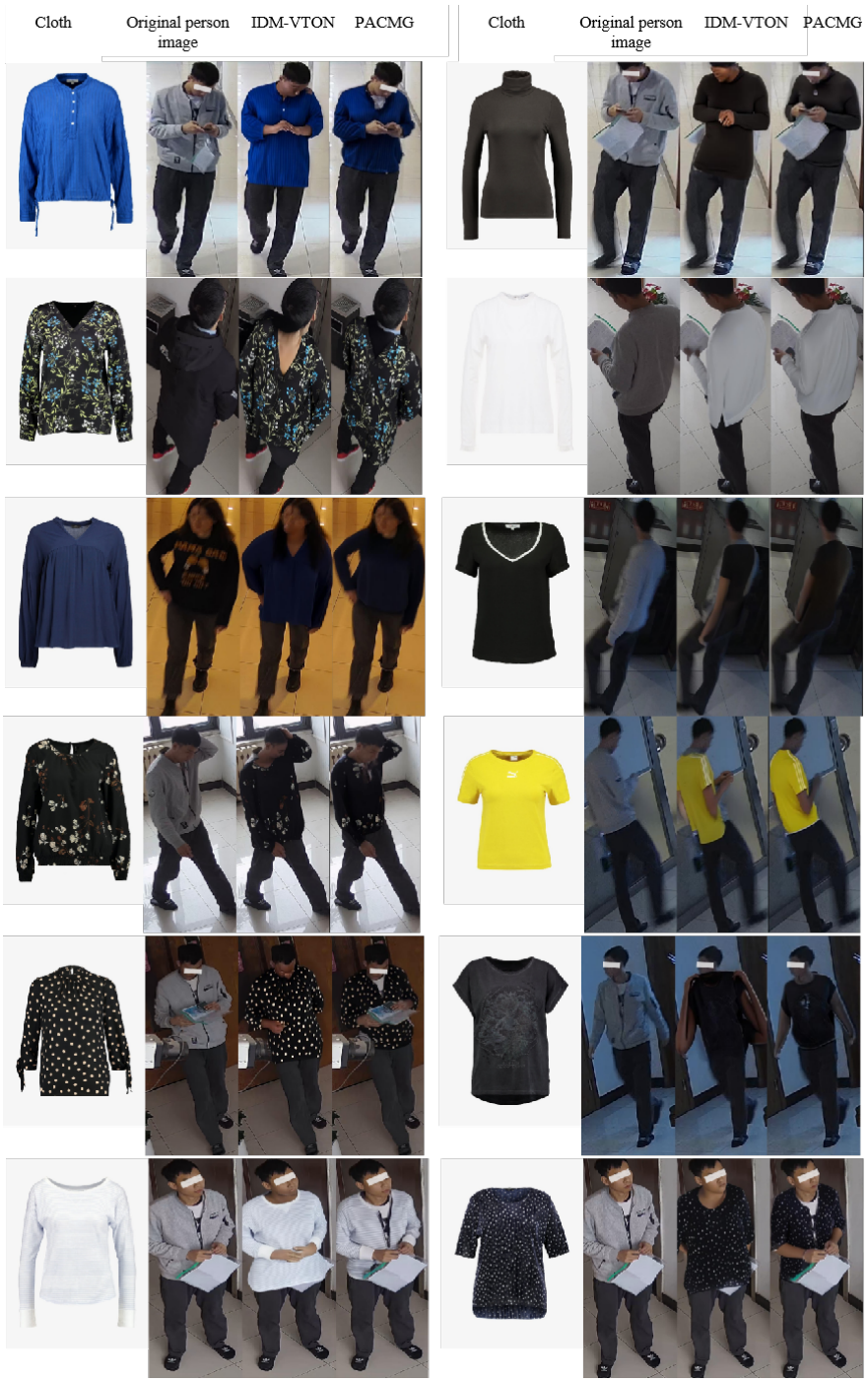


Figure 8. Comparison of Clothing Replacement Results.

The proposed method also improves the success rate of clothing replacement. A total of 2,107 person images are randomly paired with garments from the VITON-HD dataset, producing 6,371 person–garment pairs. Each person image is matched with an average of three garments. Clothing replacement is performed using both the original pipeline and the proposed method, generating 6,371 images for each. A human evaluation is then conducted to remove low-quality results, including images with facial distortion or abnormal body structure. As reported in Table 2, the original method produces 3,313 valid images, whereas the proposed method yields 4,690 valid results. These results demonstrate a substantially higher success rate and a larger number of usable augmented samples.

 doi:10.20944/preprints202601.1873.v1

Table 2. Comparison of Valid Clothing Replacement Rates Between Two Methods

IDM-VTON			PACMG			Metric
Images After Replacement	Valid Images	Validity Rate	Images After Replacement	Valid Images	Validity Rate	Comparison
6371	3313	52%	6371	4690	73.61%	Improved by 21.61%

4.2.2. Quantitative Evaluation of Image Quality

To evaluate the quality of the generated images before and after the algorithm improvement, this paper uses FID, LPIPS, and SSIM metrics for quantitative analysis. LPIPS and SSIM calculate the similarity between the original image and the clothing replacement image, respectively. Lower LPIPS values and higher SSIM values both indicate higher perceptual and structural similarity. FID assesses the difference between the generated image and the real image from the perspective of data distribution. Lower FID values indicate better matching. The results are summarized in Table 3. The proposed method outperforms the original process on all metrics, demonstrating its effectiveness in improving clothing replacement quality in surveillance scenarios.

Table 3. Quantitative Evaluation of Clothing Replacement

Method	LPIPS ↓	SSIM ↑	FID ↓
IDM-VTON	0.24	0.78	46.17
PACMG	**0.19**	**0.82**	**36.20**

Note: ↓ indicates lower values are better, ↑ indicates higher values are better.

4.3. Validity Analysis of Augmented Data

When clothing replacement is used for data augmentation in person Re-ID, two fundamental principles must be satisfied: identity consistency and distribution consistency. Identity consistency requires that the person identity remain unchanged after clothing replacement. Distribution consistency requires that the augmented data follow the same distribution as the original dataset without introducing systematic bias. This subsection evaluates the validity of the augmented data from these two aspects.

4.3.1. Identity Consistency Analysis Before and After Clothing Replacement

t-Distributed Stochastic Neighbor Embedding (t-SNE) is employed to visualize high-dimensional feature representations in a low-dimensional space. To examine whether identity information is preserved after clothing replacement, t-SNE is used to compare feature distributions of person images before and after augmentation.

Figure 9 shows the feature embeddings of real images and clothing-replaced images for the same identity. Dark-colored points represent real images, while light-colored points denote generated images. Samples belonging to the same identity share the same color.

As shown in the Figure 9, features of clothing-replaced images are largely distributed within the neighborhood of the corresponding real identity clusters. This indicates that identity-related characteristics are well preserved after augmentation. Although some generated samples exhibit increased dispersion and partial overlap with other identities in local regions, their distribution centers remain aligned with those of the original samples. No evident identity drift is observed. These results confirm that the proposed clothing replacement method maintains identity consistency while introducing appearance variations.

 doi:10.20944/preprints202601.1873.v1

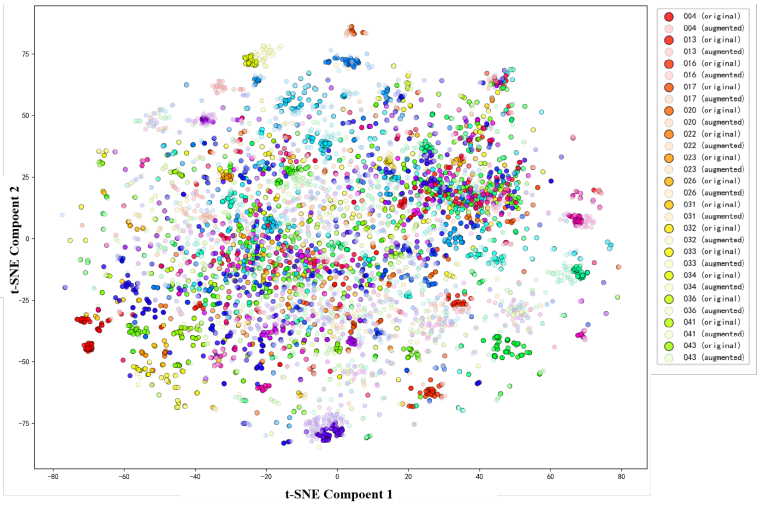


Figure 9. t-SNE visualization of identity clustering.

4.3.2. Data Distribution Consistency Analysis

This subsection evaluates the consistency between clothing-replaced images and real images from three perspectives: pixel-level statistics, perceptual similarity, and overall distribution characteristics.

Pixel-Level Statistical Analysis

As a first step, both real images and clothing-replaced images are first converted to grayscale. Low-level consistency is then examined in terms of brightness, grayscale distribution, and texture features.

- **Brightness Distribution** The results are shown in Figure 10(**a**). As shown in the figure, the brightness distribution before and after changing clothes is highly consistent, with the median being almost identical, only slightly shifted upwards. This indicates that changing clothes did not introduce a significant brightness deviation.
- **Grayscale Distribution Difference** The L2 distance was used to measure the distribution difference between the grayscale histograms of the real image and the image after the clothing change. The corresponding results are shown in Figure 10(**b**). As shown in the figure, the histogram difference exhibits a unimodal distribution, concentrated within a relatively narrow range. This indicates that the overall grayscale distribution before and after the garments change is highly consistent.
- **Texture Feature Analysis** As illustrated in Figure 10(**c**), texture differences are concentrated around zero and exhibit unimodal distributions with long tails, where extreme values account for only a small proportion of samples. These results indicate that the statistical texture properties of the clothing-replaced images are largely consistent with those of the real images.

Overall, at the pixel level, the clothing replacement process does not introduce systematic distribution shifts or texture degradation. The generated images maintain a high degree of consistency with the original images in terms of low-level statistical features.

 doi:10.20944/preprints202601.1873.v1

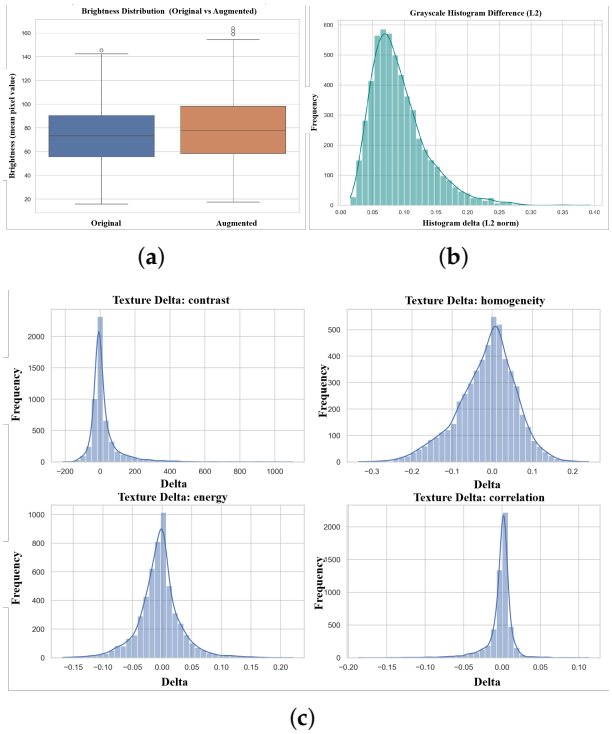


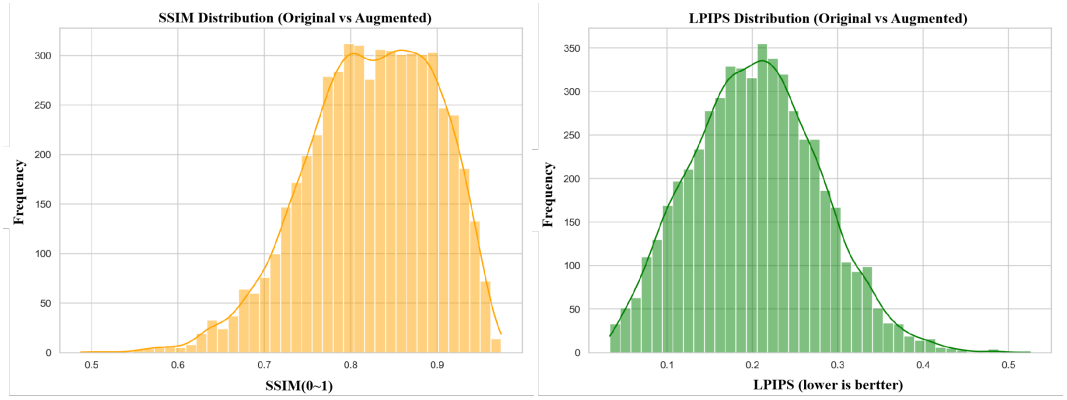
Figure 10. Pixel-Level Differences Between Clothing-Replaced and Real Images: (**a**) Brightness Comparison (before enhancement vs. after enhancement). (**b**) Grayscale Comparison (before enhancement vs. after enhancement). (**c**) Texture Comparison (before enhancement vs. after enhancement), including: Contrast, Homogeneity, Energy, and Correlation metrics.

Perceptual Similarity Analysis

This section uses SSIM and LPIPS to evaluate the structural and perceptual differences between real images and their corresponding images in different outfits. These two indices are calculated for each pair of images, and the resulting score distribution is statistically analyzed.

- **SSIM Distribution** As shown in Figure 11(**a**), SSIM values are mainly concentrated in the range of 0.75–0.92 and exhibit a unimodal distribution, with the peak located between 0.80 and 0.90. The overall distribution is smooth and stable. Since SSIM measures local structural consistency, these results indicate that the clothing-replaced images largely preserve the geometric structure and texture patterns of the original images. At the same time, SSIM values are not clustered at extremely high levels (e.g., >0.95), suggesting that the generated images are not simple copies. Instead, structural consistency is maintained while appearance variations are introduced, which is desirable for data augmentation.
- **LPIPS Distribution** As illustrated in Figure 11(**b**), LPIPS values follow a unimodal, right-skewed distribution. Most image pairs have LPIPS values below 0.3, with a small number of samples forming a long tail extending to 0.4–0.5. Because LPIPS measures perceptual dissimilarity, lower values indicate higher perceptual similarity. The observed distribution therefore suggests that the clothing-replaced images remain perceptually consistent with the original images. Meanwhile, the absence of extremely low values indicates that meaningful appearance changes are introduced. This balance between consistency and variation is well aligned with the objective of data augmentation.

 doi:10.20944/preprints202601.1873.v1



(**a**) (**b**)

Figure 11. Perceptual-Level Differences Between Clothing-Replaced and Real Images: (**a**) SSIM. (**b**) LPIPS.

Feature Distribution-Level Analysis

In this subsection, FID is used to measure the distributional discrepancy between clothing-replaced images and real images in deep feature space. The computed FID score between generated and real images is 36.2. In general practice, an FID value below 50 is regarded as indicative of high-quality image generation. The results indicate that the distribution of high-level semantic features in augmented images is close to that in real images, with a reasonable deviation, which aligns with the goal of increasing data diversity.

Comprehensive analysis shows that the proposed clothing-based data augmentation method not only increases data diversity but also maintains consistency with real images in terms of pixel-level statistics, perceptual similarity, and high-level feature distributions. The results show that a good balance is achieved between dataset enrichment, identity preservation, and distribution stability, which is crucial for robust person re-identification.

4.4. Evaluation of Person Re-ID with Augmented Data

Based on the validity analysis of the augmented data, this subsection evaluates the impact of clothing-based augmentations on the performance of person Re-ID. Comparative experiments are conducted using a person Re-ID model with a ResNet-50 backbone. All experiments are conducted on a self-collected dataset containing 159 identities and 43,742 images of person at different resolutions. The training set contained 131 identities and 39,150 images, while the test set contained 28 identities and 4,592 images. The median number of images per identity in the training set was 185.

To alleviate data imbalance while controlling the computational cost, the tiered augmentation strategy described in Section 3.3 is adopted. In our experiments, the thresholds $T_1 = 30$ and $T_2 = 60$ are empirically determined based on dataset statistics. According to this strategy, 60 identities are selected for augmentation. Specifically, identities with fewer than 30 images are augmented to the median; those with 30–59 images are augmented to $0.8\times$ the median; those with 60 images but still below the median are augmented to $0.9\times$ the median; and identities already exceeding the median are not augmented. Detailed statistics are reported in Table 4. After augmentation, the number of training images increases from 39,150 to 44,508.

Table 4. Data Augmentation Strategy

	Data Size Category		
Metric	**<30**	**30–60**	**60–Medium**
Number of IDs	20	24	16
Images Before Aug.	402	1,027	1,234
Images to Aug.	2,030	1,859	1,469

 doi:10.20944/preprints202601.1873.v1

To ensure a fair comparison, an additional control setting is introduced in which data augmentation is performed by simple image duplication, using the same identities and augmentation quantities as in the clothing-based augmentation. Three experiments are conducted under identical hardware and software configurations, with the random seed fixed across all runs:

- Experiment 1: Training on the original dataset without augmentation (Model 1);
- Experiment 2: Training on the dataset augmented using the proposed clothing replacement method (Model 2);
- Experiment 3: Training on the dataset augmented by image duplication only (Model 3).

4.4.1. Model Evaluation Results

Three ResNet-50–based person Re-ID models are obtained from the experiments described above. All models are evaluated on the same test set using Rank-1, Rank-5, Rank-10, and mAP. The quantitative results are reported in Table 5.

Table 5. Model Evaluation Results

	Number of images	R1	R5	R10	mAP
Experiment 1	39,150	29.73%	54.64%	67.12%	20.75%
Experiment 2	44,508	**32.58%**	**58.47%**	67.18%	23.08%
Experiment 3	44,508	31.77%	53.99%	64.46%	**24.26%**

As shown in Table 5, compared with the model trained on the unaugmented dataset (Experiment 1), the model trained with clothing replacement–based augmentation (Experiment 2) achieves consistent improvements across all evaluation metrics. In particular, for recall-oriented metrics such as Rank-1 and Rank-5, the gains obtained by clothing-based augmentation are substantially larger than those achieved by duplication-based augmentation (Experiment 3). The results demonstrate that clothing replacement–based augmentation could effectively enforce the model learning more discriminative person features, thereby improving the performance of person Re-ID.

4.4.2. Visualization of Performance Improvements

To further demonstrate the enhancement effect of the proposed method, three groups of retrieval results are presented in Figure 12. In each group, the first row shows the model trained on non-augmented data (Experiment 1), while the second row displays the retrieval results of model trained on clothing replacement augmented data (Experiment 2). As can be seen from the Figure 12, the number of mismatched persons among the top ten retrieval results in the second row of each group is decreased. This indicates that data obtained through virtual try-on augmentation could effectively enhance the model's performance.

 doi:10.20944/preprints202601.1873.v1

Figure 12. Comparison of Retrieval Results Before and After Virtual Try-On–Based Augmentation.

5. Conclusions

This paper introduces and refines a virtual try-on-based data augmentation method designed for person re-identification in safety-critical surveillance scenarios. The method enhances model robustness against sudden appearance changes, which is especially crucial in emergency evacuation

 doi:10.20944/preprints202601.1873.v1

situations, as reliably identifying key individuals enables timely guidance and risk mitigation. The main contributions of this work can be summarized as follows:

1. A virtual try-on–based data augmentation method for person re-identification is proposed. The pipeline is adapted to surveillance scenarios through improved clothing mask generation. The method preserves identity while enhancing visual realism and appearance diversity, enabling effective training data expansion.
2. The effectiveness of clothing-based augmentation is systematically evaluated from multiple perspectives. Comparisons with the original try-on pipeline, together with t-SNE visualization and pixel-level, perceptual-level, and distribution-level analyses, show that the method preserves identity consistency and avoids systematic distribution shifts.
3. Performance gains are verified through controlled experiments. Three comparative experiments using a ResNet-50–based model demonstrate that clothing replacement–based augmentation consistently outperforms both unaugmented training and duplication-based augmentation.

The proposed augmentation method demonstrates strong flexibility and extensibility. Clothing masks are generated via textual prompts, which reduces data preparation complexity. The prompt-based segmentation strategy enables flexible extraction of garments and accessories. It supports diverse scenarios such as lower-garment and accessory replacement, as well as targeted replacement for multiple persons in the same scene.

However, the approach has limitations. Prompt-generated masks are sensitive to color similarity. When target garments share similar colors with adjacent regions, such as hair or nearby clothing, mask accuracy may degrade, affecting replacement quality.

This work represents an initial exploration of applying virtual try-on to data augmentation for person re-identification. Future studies may enable customized replacement of arbitrary garments and accessories to produce richer datasets. With advances in generative models and virtual try-on techniques, more stable and unified frameworks for diverse clothing categories are expected. These developments will help address data scarcity and provide stronger support for Robust Person Re-ID model in emergency surveillance scenarios.

Author Contributions: Pei Wang: Conceptualization, Methodology, Investigation, Writing—original draft. Jiaming Liu: Formal analysis, Visualization, Writing—review & editing, Validation. Yuyao Cao: Writing—review & editing, Validation. Hui Zhang: Conceptualization, Methodology, Writing—review & editing, Supervision, Project administration, Validation. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (Grant No. 2024YFC3017000) and the National Natural Science Foundation of China (Grants No. 72334003 and No. 72521001).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Patient consent was waived. Public datasets were pre-anonymized by their providers. For private datasets, rigorous de-identification was performed through facial and ocular region obscuration to prevent identity recognition.

Data Availability Statement: Patient consent was waived. Public datasets were pre-anonymized by providers. Private datasets underwent facial and ocular region obscuration to ensure complete de-identification.

Data Availability Statement: This study utilizes the following publicly available datasets:

- VITON-HD: https://github.com/shadow2496/VITON-HD
- ICFG-PEDES: https://github.com/zifyloo/SSAN

Privately collected image data are not publicly shared to protect individual privacy. De-identified versions (with all facial features obscured) are available from the corresponding author upon reasonable request for research validation.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments and suggestions.

 doi:10.20944/preprints202601.1873.v1

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

Abbreviations

The following abbreviations are used in this manuscript:

Re-ID	person Re-ID
LTCC	Long-Term Clothing Changing person Re-ID
PRCC	person Re-ID under Clothing Change
GANs	Generative Adversarial Networks
IDM-VTON	Image Diffusion Model for Virtual Try-On
HBDset	Human Behavior Detection Dataset
SAM	Segment Anything Model
FID	Fréchet Inception Distance
LPIPS	Learned Perceptual Image Patch Similarity
SSIM	Structural Similarity Index Measure
t-SNE	t-Distributed Stochastic Neighbor Embedding
mAP	mean Average Precision

References

1. Sun, Z.; Wang, X.; Zhang, Y.; Song, Y.; Zhao, J.; Xu, J.; Yan, W.; Lv, C.. A comprehensive review of pedestrian re-identification based on deep learning. *Complex & Intelligent Systems* **2024**, *10*(2), 1733–1768.
2. Behera, N.K.S.; Sa, P.K.; Muhammad, K.; Bakshi, S.. Large-scale person Re-ID for crowd monitoring in emergency. *IEEE Transactions on Automation Science and Engineering* **2023**.
3. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.. Deep learning for person Re-ID: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence* **2021**, *44*(6), 2872–2893.
4. Yogameena, B.; Nagananthini, C.. Computer vision based crowd disaster avoidance system: A survey. *International journal of disaster risk reduction* **2017**, *22*, 95–129.
5. Ding, Y.; Chen, X.; Wang, Z.; Zhang, Y.; Huang, X.. Human behaviour detection dataset (HBDset) using computer vision for evacuation safety and emergency management. *Journal of Safety Science and Resilience* **2024**, *5*(3), 355–364.
6. Ding, Y.; Chen, X.; Zhang, Y.; Huang, X.. Smart building evacuation by tracking multi-camera network and explainable Re-identification model. *Engineering Applications of Artificial Intelligence* **2025**, *148*, 110394.
7. Wang, P.; Liu, J.; Peng, Y.; Zhang, H.. Intelligent Building Fire Evacuation Indication System Based on Edge Computing Devices. *2024 8th Asian Conference on Artificial Intelligence Technology (ACAIT)* **2024**, 1002–1007.
8. Ding, Z.; Ding, C.; Shao, Z.; Tao, D.. Semantically self-aligned network for text-to-image part-aware person Re-ID. *arXiv preprint arXiv:2107.12666* **2021**.
9. Sarker, P. K.; Zhao, Q.; Uddin, M. K. Transformer-based person Re-ID: a comprehensive review. *IEEE Trans. Intell. Veh.* **2024**, *9*(7), 5222–5239.
10. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based object re-identification. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021; pp. 15013–15022.
11. Ni, H.; Li, Y.; Gao, L.; Shen, H. T.; Song, J. Part-aware transformer for generalizable person Re-ID. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023; pp. 11280–11289.
12. Qian, X.; Wang, W.; Zhang, L.; Zhu, F.; Fu, Y.; Xiang, T.; Jiang, Y.; Xue, X.. Long-term cloth-changing person Re-ID. *Proceedings of the Asian conference on computer vision* **2020**.
13. Xu, P.; Zhu, X.. Deepchange: A long-term person Re-ID benchmark with clothes change. *Proceedings of the IEEE/CVF International Conference on Computer Vision* **2023**, 11196–11205.
14. Nguyen, V.D.; Mantini, P.; Shah, S.K.. Occlusion-aware appearance and shape learning for occluded cloth-changing person Re-ID. *Pattern Analysis and Applications* **2025**, *28*(2), 1–17.
15. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y.. Generative adversarial networks. *Communications of the ACM* **2020**, *63*(11), 139–144.
16. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B.. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* **2022**, 10684–10695.

 doi:10.20944/preprints202601.1873.v1

17. Choi, Y.; Kwak, S.; Lee, K.; Choi, H.; Shin, J.. Improving diffusion models for authentic virtual try-on in the wild. *European Conference on Computer Vision* **2024**, 206–235.
18. Morelli, D.; Baldrati, A.; Cartella, G.; Cornia, M.; Bertini, M.; Cucchiara, R.. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. *Proceedings of the 31st ACM international conference on multimedia* **2023**, 8580–8589.
19. Chong, Z.; Dong, X.; Li, H.; Zhang, S.; Zhang, W.; Zhang, X.; Zhao, H.; Jiang, D.; Liang, X.. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886* **2024**.
20. Xu, Y.; Gu, T.; Chen, W.; Chen, A.. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *Proceedings of the AAAI Conference on Artificial Intelligence* **2025**, *39*(9), 8996–9004.
21. Khalid, W.; Liu, B.; Waqas, M.. Clothmix: A Cloth Augmentation Strategy for Cloth-Changing person Re-ID. *2024 IEEE International Conference on Multimedia and Expo (ICME)* **2024**, 1–6.
22. Uc-Cetina, V.; Alvarez-Gonzalez, L.; Martin-Gonzalez, A.. A review on generative adversarial networks for data augmentation in person Re-ID systems. *arXiv preprint arXiv:2302.09119* **2023**.
23. Nguyen, V.D.; Mantini, P.; Shah, S.K.. Contrastive clothing and pose generation for cloth-changing person Re-ID. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2024**, 7541–7549.
24. Jiang, L.; Zhang, C.; Wu, L.; Li, Z.; Wang, Z.; Wei, C.. Joint feature augmentation and posture label for cloth-changing person Re-ID. *Multimedia Systems* **2025**, *31*(2), 103.
25. Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; Kautz, J.. Joint discriminative and generative learning for person Re-ID. *proceedings of the IEEE/CVF conference on computer vision and pattern recognition* **2019**, 2138–2147.
26. Yu, Z.; Zhao, Y.; Hong, B.; Jin, Z.; Huang, J.; Cai, D.; Hua, X.. Apparel-invariant feature learning for person Re-ID. *IEEE Transactions on Multimedia* **2021**, *24*, 4482–4492.
27. Wei, L.; Zhang, S.; Gao, W.; Tian, Q.. Person transfer gan to bridge domain gap for person Re-ID. *Proceedings of the IEEE conference on computer vision and pattern recognition* **2018**, 79–88.
28. Cui, A.; Mahajan, J.; Shah, V.; Gomathinayagam, P.; Liu, C.; Lazebnik, S.. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2024**, 8235–8239.
29. Zhang, X.; Song, D.; Zhan, P.; Chang, T.; Zeng, J.; Chen, Q.; Luo, W.; Liu, A.. Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. *Proceedings of the Computer Vision and Pattern Recognition Conference* **2025**, 26399–26408.
30. Subramaniam, R.R.; Dhamini, M.; Srija, M.; Vaishnavi, M.; Vidyadhari, M.. Enhancing E-Commerce with Virtual Try-On Using Stable VITON. *2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE)* **2025**, 1–5.
31. Pathak, S.; Kaushik, V.; Lall, B.. GraVITON: Graph based garment warping with attention guided inversion for Virtual-tryon. *Proceedings of the Asian Conference on Computer Vision* **2024**, 573–588.
32. Gou, J.; Sun, S.; Zhang, J.; Si, J.; Qian, C.; Zhang, L.. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. *Proceedings of the 31st ACM International Conference on Multimedia* **2023**, 7599–7607.
33. Lee, S.; Gu, G.; Park, S.; Choi, S.; Choo, J.. High-resolution virtual try-on with misalignment and occlusion-handled conditions. *European Conference on Computer Vision* **2022**, 204–219.
34. Shim, S.; Chung, J.; Heo, J.. Towards squeezing-averse virtual try-on via sequential deformation. *Proceedings of the AAAI Conference on Artificial Intelligence* **2024**, *38*(5), 4856–4863.
35. Velioglu, R.; Bevandic, P.; Chan, R.; Hammer, B.. Enhancing Person-to-Person Virtual Try-On with Multi-Garment Virtual Try-Off. *arXiv preprint arXiv:2504.13078* **2025**.
36. Song, D.; Zhang, X.; Zhou, J.; Nie, W.; Tong, R.; Kankanhalli, M.; Liu, A.. Image-based virtual try-on: A survey. *International Journal of Computer Vision* **2025**, *133*(5), 2692–2720.
37. Li, N.; Liu, Q.; Singh, K.K.; Wang, Y.; Zhang, J.; Plummer, B.A.; Lin, Z.. Unihuman: A unified model for editing human images in the wild. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* **2024**, 2039–2048.
38. Feng, Y.; Zhang, L.; Cao, H.; Chen, Y.; Feng, X.; Cao, J.; Wu, Y.; Wang, B.. Omnitry: Virtual try-on anything without masks. *arXiv preprint arXiv:2508.13632* **2025**.
39. Wan, Z.; Hu, D.; Cheng, W.; Chen, T.; Wang, Z.; Liu, F.; Liu, T.; Gong, M.; others. Mf-viton: High-fidelity mask-free virtual try-on with minimal input. *arXiv preprint arXiv:2503.08650* **2025**.
40. Issenhuth, T.; Mary, J.; Calauzenes, C.. Do not mask what you do not need to mask: a parser-free virtual try-on. *European Conference on Computer Vision* **2020**, 619–635.

 doi:10.20944/preprints202601.1873.v1

41. Niu, Y.; Wu, L.; Yi, D.; Peng, J.; Jiang, N.; Wu, H.; Wang, J.. Anydesign: Versatile area fashion editing via mask-free diffusion. *arXiv preprint arXiv:2408.11553* **2024**.
42. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; others. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *European conference on computer vision* **2024**, 38–55.
43. Morelli, D.; Fincato, M.; Cornia, M.; Landi, F.; Cesari, F.; Cucchiara, R.. Dress code: High-resolution multi-category virtual try-on. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* **2022**, 2231–2235.
44. Choi, S.; Park, S.; Lee, M.; Choo, J.. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* **2021**, 14131–14140.
45. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. Segment anything. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023; pp. 4015–4026.
46. Li, P.; Xu, Y.; Wei, Y.; Yang, Y.. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *44*(6), 3260–3271.
47. Cao, Z.; Simon, T.; Wei, S.; Sheikh, Y.. Realtime multi-person 2d pose estimation using part affinity fields. *Proceedings of the IEEE conference on computer vision and pattern recognition* **2017**, 7291–7299.
48. Güler, R.A.; Neverova, N.; Kokkinos, I.. Densepose: Dense human pose estimation in the wild. *Proceedings of the IEEE conference on computer vision and pattern recognition* **2018**, 7297–7306.