# OmniPerson: Unified Identity-Preserving Pedestrian Generation

Changxiao Ma[1,2], Chao Yuan[1], Xincheng Shi[1], Yuzhuo Ma[1], Yongfei Zhang[1,2,†]
Longkun Zhou[2], Yujia Zhang[1], Shangze Li[1], Yifan Xu[1]
[1] Beihang University [2] Pengcheng Laboratory
*Codes:* https://github.com/maxiaoxsi/OmniPerson

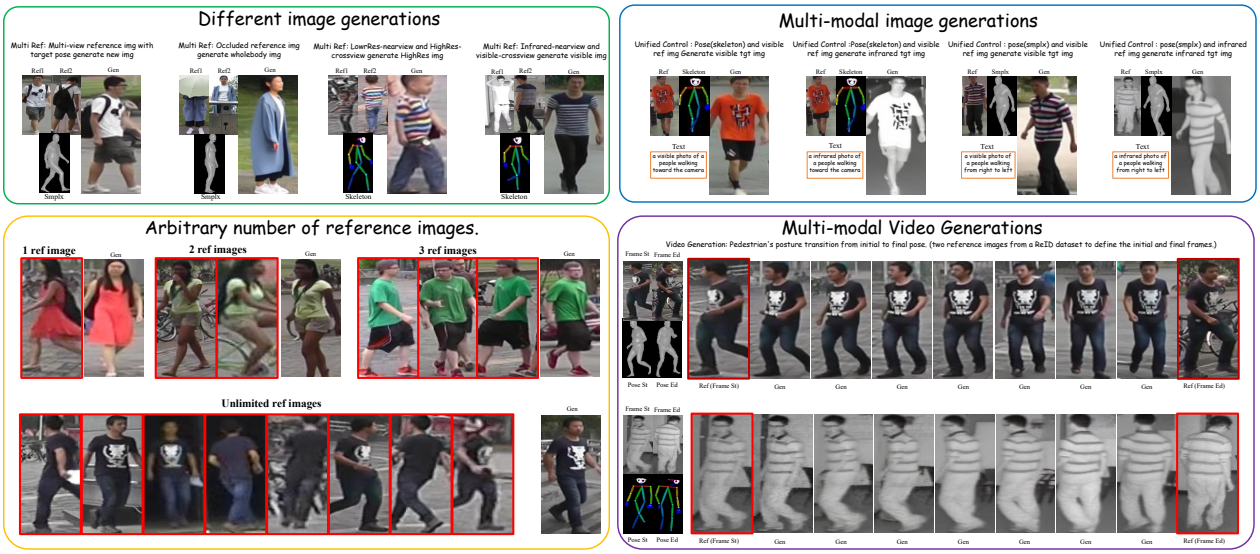


Figure 1. The Muti-tasks generation results of **OmniPerson**: Controllable Pedestrian Generation from Diverse References.

## Abstract

*Person re-identification (ReID) suffers from a lack of large-scale high-quality training data due to challenges in data privacy and annotation costs. While previous approaches have explored pedestrian generation for data augmentation, they often fail to ensure identity consistency and suffer from insufficient controllability, thereby limiting their effectiveness in dataset augmentation. To address this, We introduce **OmniPerson**, the first unified identity-preserving pedestrian generation pipeline for visible/infrared image/video ReID tasks. Our contributions are threefold: ① We proposed **OmniPerson**, a **unified generation model**, offering holistic and fine-grained control over all key pedestrian attributes. Supporting **RGB/IR modality image/video generation** with any number of reference images, two kinds of person poses, and text. Also including **RGB-to-IR transfer** and **image super-resolution** abilities. ② We designed **Multi-Refer Fuser** for robust **identity preservation** with any number of reference images as input, making **OmniPerson** could distill a unified identity from a set of multi-view reference images, ensuring our generated pedestrians achieve high-fidelity pedestrian generation. ③ We introduce **PersonSyn**, the first large-scale dataset for **multi-reference, controllable** pedestrian generation, and present its automated curation pipeline which transforms public, ID-only ReID benchmarks into a richly annotated resource with the **dense, multi-modal supervision** required for this task. Experimental results demonstrate that OmniPerson achieves SoTA in pedestrian generation, excelling in both visual fidelity and identity consistency. Furthermore, augmenting existing datasets with our generated data consistently improves the performance of ReID models. We will open-source the full codebase, pre-trained model, and the PersonSyn dataset.*

## 1. Introduction

Person re-identification (ReID) aims to retrieve and track target individuals across non-overlapping camera views based on appearance features (e.g. clothing and body

shape). In recent years, the field has been dominated by deep learning models[1] that have achieved remarkable performance on established benchmarks[16, 23].

The success of these data-driven methods is, however, critically dependent on the availability of large-scale, diverse, and meticulously annotated datasets. This reliance on massive datasets constitutes a major real-world bottleneck, fraught with challenges including prohibitive annotation costs, stringent data privacy regulations, and the difficulty of capturing the long tail of variations in appearance, pose, and environmental conditions[37]. This data scarcity severely limits the robustness and generalizability of even the SoTA ReID models in real-world deployments[9], motivating the exploration of alternative data sources.

Generative data augmentation has emerged as a powerful strategy to overcome the challenges of data scarcity and limited diversity in ReID. The value of this approach is well-established, as extensive research has demonstrated that augmenting training sets with high-quality synthetic data significantly enhances both the discriminative power [33] and generalization capability [57] of ReID models. The technology enabling this has progressed rapidly, from early methods using classical image processing [60] and virtual engines [57], to the first wave of generative approaches based on GANs [43]. More recently, the advent of large-scale diffusion models, such as Stable Diffusion [38], has marked a paradigm shift in image synthesis. Capitalizing on their unprecedented realism and control [10], the latest works now focus on applying these diffusion-based models specifically for high-fidelity **person synthesis** [8, 29].

However, despite the strong generative capabilities of these state-of-the-art **person generation** models, applying them directly to the specialized task of ReID data augmentation reveals two critical limitations.

(1) First, they exhibit a critical failure in **high-fidelity identity preservation**, especially when generating subjects under drastic variations in pose and viewpoint. The core challenge in ReID data augmentation is creating these novel views, but this is precisely where existing methods, which typically rely on a single reference image for appearance guidance [8, 29, 50], are fundamentally flawed. As illustrated in Figure 2, a single view provides an incomplete and often biased representation, failing to capture details visible only from other angles (e.g., the back of a shirt or a logo on a sleeve). Consequently, when tasked with rendering the person from a novel perspective, the model is compelled to hallucinate missing visual information, resulting in significant inconsistencies and "identity drift." This introduces conflicting signals into the ReID training data; instead of learning a robust identity signature, the ReID model is fed ambiguous information that can corrupt its feature space and degrade performance.

(2) Second, they lack a unified and disentangled control

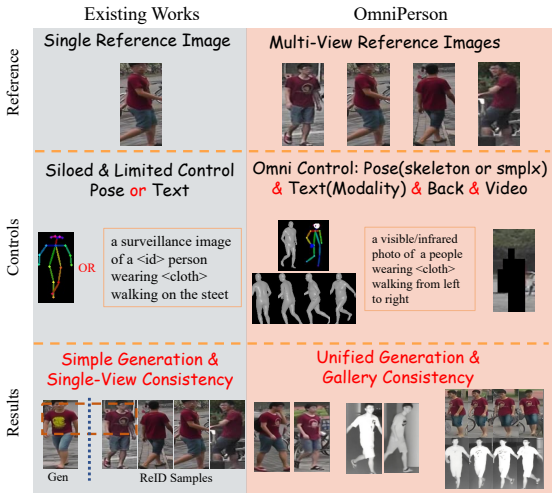


Figure 2. Exiting works vs OmniPerson.

mechanism. Existing methods operate in silos: text-driven approaches struggle to faithfully condition on a specific visual identity, often leading to inconsistent results.[40] Conversely, methods conditioned on a visual reference are often pose-driven, offering little to no explicit control over crucial attributes like background or modality. This fragmented control scheme prevents the targeted synthesis of challenging, long-tail scenarios (e.g., a specific person in a rare pose against a novel background, captured in the infrared spectrum) that are crucial for systematically bolstering the robustness of ReID models.

To overcome these fundamental limitations of identity drift and fragmented control, we introduce **OmniPerson**, a unified generation pipeline designed specifically for identity-preserving pedestrian data synthesis, and **PersonSyn**, a detailed reconstructed dataset for pedestrian generation. Our main contributions are four-fold:

- **A Unified Identity-Preserving Pedestrian Generation model: OmniPerson.** Supporting **RGB/IR modality image/video generation** with any number of reference images, two kinds of person poses, and text. Also including **RGB-to-IR transfer** and **image super-resolution** abilities.
- **Multi-Refer Fuser for OmniPerson.** It achieve the robust **identity preservation** with any number of reference images as input, making **OmniPerson** could distill a unified identity from a set of multi-view reference images, ensuring our generated pedestrians achieve high-fidelity pedestrian generation.
- **PersonSyn Dataset.** To facilitate research in multi-reference, controllable pedestrian generation, we reorganized the public pedestrian datasets and obtained **PersonSyn** dataset with the comprehensive annotations for pedestrian generation, including **2D/3D poses, disentan-**

**gled background plates, classified viewpoints, textual attributes, and reference image sets.**

- Extensive experiments demonstrate that OmniPerson achieves SOTA performance in both generation quality and identity consistency. Furthermore, we validate on downstream tasks and achieve SOTA performance.

## 2. Related Works

### 2.1. Person Re-Identification

Person re-identification has become a cornerstone task in computer vision, aiming to associate pedestrian identities across non-overlapping camera views and varying time instances. Recent advances in ReID have been largely driven by deep learning, with convolutional neural networks (CNNs) [20] and vision transformers (ViTs) [1] proving particularly effective in boosting recognition accuracy. State-of-the-art approaches typically integrate feature extraction with metric learning frameworks [21, 25, 26, 42, 49, 51–53], enabling the learning of discriminative representations that remain robust to challenges such as pose variations, viewpoint changes, and illumination differences.

Standard datasets such as Market-1501 [58] have been widely adopted to evaluate ReID algorithms under controlled conditions, while SYSU-MM01 [46] focuses on matching identities between visible and infrared images. Although existing methods demonstrate strong benchmark performance on these datasets, the limited number of images and identities, along with homogeneity of collection scenarios, result in ReID models lacking robustness in real-world deployments.

### 2.2. Data Augmentation for ReID

To address data scarcity in person ReID, prior works have employed Generative Adversarial Networks (GANs) [15] and virtual simulation engines[34] to synthesize pedestrian images[43, 61] for dataset augmentation. While GAN-based approaches have proven effective for ReID model enhancement through style transfer [7, 30, 59] and pose-guided generation [14, 33], virtual engine-based methods demonstrate that large-scale synthetic pedestrian data can significantly improve model generalizability [22, 45, 57]. However, the drawback of these methods is the insufficient quality and realism of the generated images, which limits their effectiveness. This fidelity gap has motivated the exploration of more powerful generative frameworks.

### 2.3. Diffusion Model

The field of image generation has witnessed advancement with diffusion models [17]. Stable Diffusion [38] enables high-quality synthesis via latent diffusion, while text conditioning [35, 39] enhances controllability. ControlNet [54] further introduces conditional constraints into diffusion architectures.

Extensive research has demonstrated that diffusion-based methods can generate high-quality pedestrian images [2, 50]. Subsequent work [29] has further validated that Stable Diffusion-based generation can effectively augment ReID datasets. However, despite significant quality improvements over earlier approaches [10], maintaining identity-consistent controllable generation remains an open challenge for diffusion-based methods.

Our proposed OmniPerson addresses these limitations through a multi-reference diffusion framework that achieves both identity-consistent generation and simultaneous modality, pose, and background control for ReID data augmentation.

## 3. The PersonSyn Dataset

Public ReID datasets lack the rich annotations needed for controllable person generation as they only provide identity supervision. We introduce PersonSyn, a large-scale dataset that transforms these weakly-labeled benchmarks by adding fine-grained annotations for pose, viewpoint, background, visual attributes, and a structured reference set. We now detail our curation pipeline, the automated process designed to generate the comprehensive annotations for the PersonSyn dataset.

**Condition Extraction** The initial Condition Extraction stage derives a rich set of multi-modal annotations from each pedestrian image. We employ OpenPose [4] to extract 2D skeletal keypoints and concurrently use SMPLest-X [3] to recover 3D human mesh (SMPL-X [32]) parameters. These SMPL-X parameters are then leveraged for three key tasks: to render a 3D model, to compute the pedestrian's orientation unit vector, and to enable background extraction. In parallel, semantic attributes are extracted using the Qwen-VL vision-language model[44]. **A comprehensive list of these attributes is detailed in Supplementary materials.**

**Data Cleaning** To ensure annotation quality, we apply a strict filtering protocol, removing any samples that meet the following criteria:(1) Poor overall visual quality, (2) Low confidence scores from the pose estimators. (3) An insufficient number of detected keypoints (below a threshold). (4) Significant misalignment between the 2D keypoints and the 3D mesh.

**Classification and Integration** In this stage, we structure the dataset by first partitioning all intra-identity references into subsets based on their orientation, camera ID, and key attributes (e.g., backpack status). We then pre-compute two

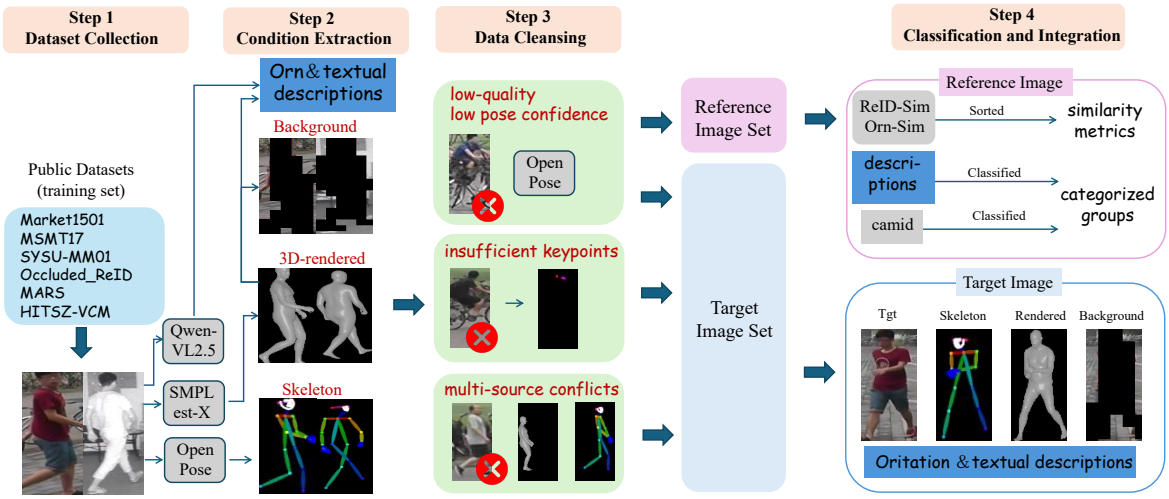


Figure 3. The pipeline of PersonSyn dataset generating with all training set of public datasets.

metrics for each potential reference-target pair: a CLIP-ReID feature similarity score and a viewpoint score derived from the dot product of their orientation unit vectors.

Crucially, the goal is not to create one fixed reference set. Instead, this process yields a flexible framework of categorized groups and rich similarity metrics, empowering researchers to design their own sampling logic. For instance, a researcher can generate a person with a backpack, using multi reference images of the same person without one and the guidance of a text prompt "person with a backpack". **More details about PersonSyn and the specific strategy we designed for training OmniPerson is detailed in the Supplementary Materials**. This entire structured collection of images, annotations, and pre-computed metrics constitutes the final PersonSyn dataset, whitch provides a comprehensive package for each image:

- **Conditional Inputs:** Paired 2D skeletons, rendered 3D meshes, and isolated backgrounds.
- **Fine-Grained Annotations:** Labels for orientation and key semantic attributes (e.g., clothing, backpack status).
- **Flexible Reference Framework:** A pre-processed database of all potential references, partitioned into subsets and scored by visual similarity and viewpoint difference.

# 4. Methods

This section details the methodology of OmniPerson, our unified generation pipeline. We begin by presenting the overall framework (Sec. 4.1). Next, we elaborate on our core contribution for identity preservation using multi-view references (Sec. 4.2). Finally, we describe the versatile conditioning mechanisms for fine-grained control (Sec. 4.3).

## 4.1. Unified Generation Framework

Our approach is built upon Latent Diffusion Model (LDM) [38], which synthesize images $x$ in an efficient latent space. LDM employs a denoising U-Net, $\epsilon_\theta$, to iteratively reverse a diffusion process, guided by a condition $c$. This process transforms a random noise vector $z_t$ into a clean latent representation $z_0$, which is then rendered into a high-resolution image by a VAE decoder$x = D(z_0)$.

While prior art in pedestrian synthesis often relies on limited guidance (typically a text prompt or a single reference image coupled with pose control). These approaches are insufficient for the structured, fine-grained control required over a pedestrian's holistic identity, pose, background, and modality.

Therefore, our key contribution is to replace this monolithic condition with a unified, multifaceted conditioning framework designed specifically for pedestrian generation. As illustrated in Figure 4, this framework is guided by a suite of distinct control signals:

1. A set of $N$ multi-view reference images for identity.
2. A target pose representation(skeleton/smplx).
3. A background scene(optional).
4. A textual descriptor specifying the modality.

The generation pipeline proceeds as follows: First, the multi-view reference images are processed by our specialized Multi-Ref Fuser Module (see Sec. 4.2) to obtain a robust identity embedding $c_{id}$. Meanwhile, the other conditioning inputs are encoded by their respective encoders to produce embeddings for pose ($c_{pose}$), background ($c_{bg}$), and text ($c_{text}$), as detailed in Sec. 4.3. These embeddings are jointly integrated into the denoising U-Net, guiding the diffusion process to transform random noise $z_t$ into a structured latent code $z_0$ that coherently aligns with all input conditions.

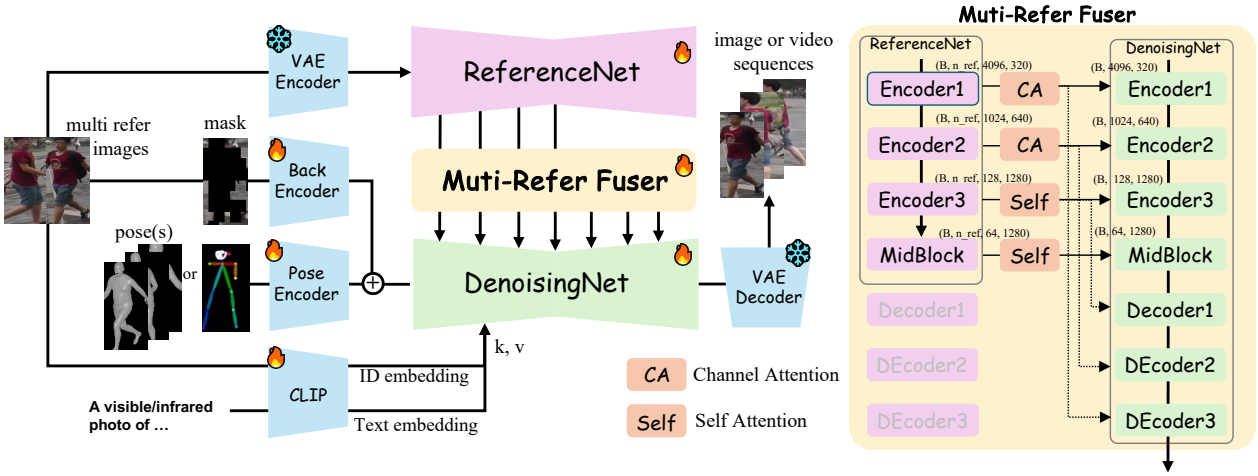


Figure 4. Overview of the **OmniPerson** framework. Our model leverages the novel Multi-Refer Fuser module to aggregate features from an arbitrary number of reference images. This approach ensures robust gallery-level identity consistency, overcoming the limitations of single-reference methods that are confined to image-level consistency. Furthermore, the framework provides fine-grained, decoupled control over (1) pose (via SMPL-X or skeleton), (2) modality, and (3) background.

### 4.2. Multi-Refer Fusion for Identity Preservation

Meaningful data augmentation for ReID requires synthesizing pedestrians across significant variations in pose and viewpoint. This task is fundamentally challenging because a single reference image offers only an incomplete, view-dependent snapshot of a person's appearance, often leading to identity shift during generation. Our approach overcomes this limitation by introducing a mechanism to distill a complete and robust identity representation from the entire set of available reference images for a person.

The fusion process begins by feeding the set of $N (N \geq 1)$ reference images through a ReferenceNet. To ensure feature space compatibility, this network shares its architecture and weights with the denoising UNet [18]. While the ReferenceNet has a full encoder-decoder structure, to balance computational efficiency and sampling quality, followed[54], we selectively extract features only from its three down-sampling encoder blocks and the middle block. Each U-Net block is composed of a self-attention layer for spatial feature extraction and a cross-attention layer for integrating semantic information. To isolate pure visual information, we disable the cross-attention mechanism and exclusively collect the output features from the self-attention layer in each block. These features serve as the multi-level pedestrian representations for fusion.

Our key insight is that features from different network depths represent different levels of abstraction and thus benefit from distinct fusion strategies. We categorize these features into two types and process them accordingly within our **Multi-Refer Fuser** module.

- **Low-Level Features**: Extracted from the early stages (down0, down1), these features represent fine-grained visual details such as texture, color, and local patterns.
- **High-Level Features**: Extracted from the deeper stages (down2, mid), these features encode more abstract, semantic information that defines the person's core identity.

**Fusion of Low-Level Features via Channel Attention**. For low-level features, the goal is to select the most salient textural and color information across the reference set. We employ a Channel Attention Fusion mechanism. For a given feature level (e.g., down0), feature maps from all $N$ reference images are concatenated. To capture global channel-wise statistics, we apply both average and max pooling along the spatial dimensions and process the resulting descriptors with a MLP to compute a shared channel-wise attention vector. This vector, passed through a sigmoid function, assigns an importance score to each channel, effectively amplifying informative features (e.g., unique clothing patterns) and suppressing less useful ones. A masked average is then performed across the reference dimension to produce a single, robustly fused low-level feature map.

**Fusion of High-Level Features via Self-Attention**. For high-level features, it is crucial to understand the contextual relationships between semantic parts across multiple views. To this end, we employ a Self-Attention Fusion mechanism. For a given deep feature level (e.g., mid), the feature tokens from all reference images are concatenated into a single sequence. Standard multi-head self-attention is then applied to this sequence, allowing each feature token to attend to every other token, regardless of its originating image. This enables the model to identify and aggregate the most representative identity-defining features. for instance, a clear frontal view of a face can inform the features of a profile

view. A average across the reference dimension yields the final fused semantic feature map.

Finally, the set of fused multi-level features from the RefFuser are integrated into the attention computation of each transformer block: denoising U-Net's self-attention module augments its key–value set with the fused reference features, enabling the model to perform cross-instance attention.

Through this mechanism, we extract and fuse multi-level identity features, which are then injected into the denoising UNet to provide a rich and comprehensive representation of the pedestrian's appearance.

### 4.3. OmniPerson pipeline

The versatility of the OmniPerson framework stems from its ability to be guided by a rich, multi-source condition vector, $c$. This condition is a composite of several distinct guidance signals that provide fine-grained control over the generated pedestrian and their environment. In this section, we detail the specific encoders and mechanisms used to process each of these signals: pose, background, textual attributes, and temporal context for video.

**Identity Guidance** While the mechanism detailed in Section 4.2 provides a rich and comprehensive representation of the person's visual identity, this alone is insufficient. The denoising UNet needs additional contextual conditioning to properly leverage these features. The goal is to generate a pedestrian that adheres to the target image's specific conditions, like its unique lighting and camera properties.

Our key insight is that standard ReID features, while potentially 'weak' for ReID task, are perfectly suited for this purpose. This is because ReID features are inherently entangled with both camera domain information and high-level pedestrian semantics—precisely the information needed to ground the synthesized image in the desired context.

Therefore, we employ a pretrained CLIP-ReID model to extract a ReID feature from a reference pedestrian (sharing the same ID and camera). This feature, used in conjunction with the text embedding, serves jointly as the key (K) and value (V) for the cross-attention layers within the denoising UNet, thereby guiding the generation of the target pedestrian.

**Pose and Background Guidance** To control the person's structure and the surrounding scene, we employ dedicated spatial encoders for pose and background.

- Pose Encoder: This module provides structural guidance. It is designed with the flexibility to process both 2D skeletal keypoints and rendered 3D mesh images. The encoder consists of four convolutional blocks with increasing channel sizes (16, 32, 64, 128), each using strided convolutions to downsample the input. This produces a high-dimensional pose embedding, $E_{pose}$
- Background Encoder: To explicitly control the environment, a Background Encoder with an identical architecture processes a desired background image, yielding a spatial embedding, $E_{bg}$, that captures the scene's characteristics.

The two spatial embeddings are first combined via element-wise addition to form a single, unified guidance map: $E_{spatial} = E_{pose} + E_{bg}$, and then added directly to the noisy latent $z_t$ before it enters the U-Net.

**Text-based Attribute and Modality Control** OmniPerson also supports semantic control via natural language. We construct a textual descriptor by combining the rich attributes annotated in our PersonSyn dataset (e.g., "a person wearing a blue jacket") with the desired modality (e.g., "a visible photo"). This prompt is then fed into the frozen text encoder of a pretrained CLIP model to generate a powerful semantic embedding, $E_{text}$. This embedding is integrated into the denoising U-Net via the standard cross-attention mechanism, allowing it to guide the overall attributes and modality of the generated image.

**Temporal Coherence for Video Synthesis** To extend OmniPerson from image to video synthesis, followed AnimateAnyone [18], we integrate a temporal coherence module into the U-Net. This module consists of temporal attention layers added to each unet block. When generating the current frame, its intermediate features serve as the query, while the corresponding features from the previously generated frame act as the key and value. To generate videos with continuous pose transformations, particularly across large viewpoint changes, we construct a sequence of intermediate poses to guide the synthesis process, as illustrated in Figure 5. **The pseudocode for this procedure is provided in the Supplementary Materials**.

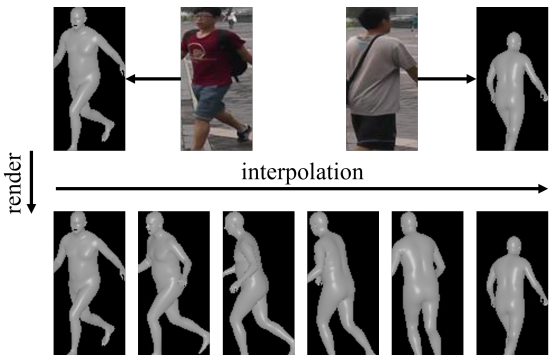


Figure 5. Pose sequence for video generation

**Training Objective** The entire OmniPerson framework is optimized end-to-end. The denoising U-Net, $\epsilon_\theta$, is trained to predict the noise added to a latent variable, conditioned on the full set of guidance signals. The objective is the standard L2 loss:

$$L = \mathbb{E}z_0, t, \epsilon, c(||\epsilon - \epsilon_\theta(z_t, t, c)||_2^2) \quad (1)$$

where $c_{id}$, $c_{pose}$, $c_{bg}$, $c_{text}$, $c_{tem}$ represents the condition of identity, pose, background, text, and temporal condition. For video clips, this loss is computed and averaged across all frames. During training, we employ classifier-free guidance by randomly dropping conditions to enhance controllability at inference time.

# 5. Experiment

## 5.1. Implementations

We train OmniPerson on PersonSyn using a single NVIDIA A100 GPU. For each training sample, we curate a reference set by selecting at least one intra-camera image and one image from each orientation set. The selection from each orientation set is probabilistic: candidates are first ranked by ReID similarity, and one is then chosen based on a rank-weighted Bernoulli sampling scheme. Additionally, to increase task difficulty, reference image with the same orientation as the target are probabilistically dropped. First, we conduct 40K training steps on image datasets with a batch size of 8 and learning rate of $1 \times 10^{-5}$. Subsequently, we switch to video training for 40K training steps with a reduced batch size of 2 (processing 8 frames per video) while maintaining the same learning rate. Finally, we fine-tune the model on image datasets for an additional 20K steps with batch size 8, lowering the learning rate to $5 \times 10^{-6}$. All input images $I \in R^{w \times h}$(we assume h $>$ w in our formulation, where h and w represent the height and width of the input image) undergo preprocessing where they are first resized to $I \in R^{(\frac{w}{h} \times 512) \times 512}$ to preserve aspect ratios, then zero-padded to $I \in R^{512 \times 512}$. **Detailed data augmentation methods are provided in the supplementary material**.

## 5.2. Generated Pedestrian Samples

Figure 1 showcases the versatile capabilities of OmniPerson. The first row demonstrates the model's core strengths in multi-reference fusion and fine-grained controllable generation. This includes complex fusion tasks, such as generating a complete body from occluded inputs or performing cross-modal (Visible-IR) synthesis. Concurrently, it displays decoupled control over pose (dictated by a skeleton or SMPL-X model) and appearance attributes (guided by textual prompts).

The subsequent two rows on the left side of the figure highlight OmniPerson's flexibility in leveraging an arbitrary number of reference images. These examples show successful synthesis using one, two, and a larger set of references, validating the robustness of our fusion mechanism.

The two rows on the right side showcase our video generation capability. Given start and end frames with a significant pose gap, OmniPerson synthesizes a smooth and temporally coherent video that interpolates the motion between them while preserving identity. **Additional generation results are provided in the supplementary material, including a qualitative comparison of our method against several baseline approaches.**.

Table 1. Generation Quality Comparison on PersonSyn

| Model | LPIPS↓ | SSIM↑ | PSNR↑ | FVD↓ |
|---|---|---|---|---|
| DSC[41] | 0.303 | 0.305 | 14.308 | - |
| PATN[61] | 0.254 | 0.282 | 14.262 | - |
| DIAF([24]) | 0.306 | 0.305 | 14.201 | - |
| FD-GAN[11] | 0.612 | 0.210 | 13.981 | - |
| DIST[36] | 0.282 | 0.281 | 14.337 | - |
| XingGAN[43] | 0.306 | 0.304 | 14.446 | - |
| SPIG[28] | 0.278 | 0.314 | 14.489 | - |
| DPTN[55] | 0.271 | 0.285 | 14.521 | - |
| AnimateAnyone[18] | 0.315 | 0.231 | 13.442 | 395.21 |
| Pose2Id[50] | 0.309 | 0.236 | 13.900 | 392.1 |
| Ours | **0.202** | **0.467** | **17.157** | **232.34** |

Table 2. Generation Quality Comparison on Market-1501

| Model | CLIP-Sim↑ | DINO-Sim↑ | ReID-Sim↑ |
|---|---|---|---|
| DPTN[55] | 0.8636 | 0.6450 | 0.9023 |
| AnimateAnyone[18] | 0.8661 | 0.6752 | 0.9131 |
| Pose2Id[50] | 0.8669 | 0.6825 | 0.9161 |
| Ours | **0.8697** | **0.8037** | **0.9577** |

## 5.3. Generation Comparison with SoTA

In this section, we compare OmniPerson with state-of-the-art (SoTA) pedestrian generation methods in terms of both generation quality and identity consistency. For each identity in PersonSyn, a target image is randomly selected, with the remaining images serving as the reference pool.

**For generation quality**, we conduct a comprehensive evaluation using a suite of standard image and video metrics. Each metric is chosen to assess a distinct aspect of quality:

- **LPIPS (Learned Perceptual Image Patch Similarity)** assesses perceptual similarity, closely aligning with human judgments of image quality.
- **SSIM and PSNR** evaluate lower-level image fidelity, with SSIM measuring structural similarity and PSNR quantifying pixel-wise reconstruction error.
- **FVD (Fréchet Video Distance)** measures the quality and temporal coherence of the generated videos by comparing their feature distributions to those of real videos.

Table 3. Ablation Study on OmniPerson Components

| Mul-Ref | ReID | Ref-Sel | LPIPS↓ | SSIM↑ | PSNR↑ |
|---|---|---|---|---|---|
| ✗ | ✗ | ✗ | 0.255 | 0.374 | 15.518 |
| ✗ | ✓ | ✗ | 0.234 | 0.420 | 16.661 |
| ✓ | ✗ | ✗ | 0.230 | 0.435 | 16.851 |
| ✓ | ✓ | ✗ | 0.210 | 0.455 | 16.912 |
| ✓ | ✓ | ✓ | 0.202 | 0.467 | 17.157 |

To ensure a fair comparison, all methods, including ours, were evaluated without explicit background control. The results clearly indicate that OmniPerson achieves state-of-the-art performance across all key image and video generation quality metrics.

**For identity consistency**, we measure the feature similarity between the generated images and their ground-truth target samples across three distinct feature spaces: CLIP, DINOv2, and ReID model(pretrained CLIP-ReID). Each metric is chosen to evaluate a different facet of similarity:

- **CLIP Similarity** captures the alignment of high-level semantic and stylistic information.
- **DINOv2 Similarity** assesses the perceptual and structural correspondence between the images.
- **ReID Similarity** directly evaluates the core objective of identity preservation.

The experimental results demonstrate that OmniPerson outperforms all existing SoTA methods across every metric, thereby exhibiting the strongest identity consistency.

### 5.4. Ablation Analysis

The results of our ablation study are presented in Table 3. We systematically analyze the contribution of three key components: Mul-Ref (our multi-reference fusion mechanism), ReID (the ReID feature guidance), and Ref-Sel (our reference selection strategy for the PersonSyn dataset).

As the results indicate, both the ReID feature guidance and the Multi-Ref fusion module independently improve the quality of the generated images. Furthermore, the experiments demonstrate that when both components are active, employing a well-designed reference selection strategy (Ref-Sel) provides an additional and significant performance boost.

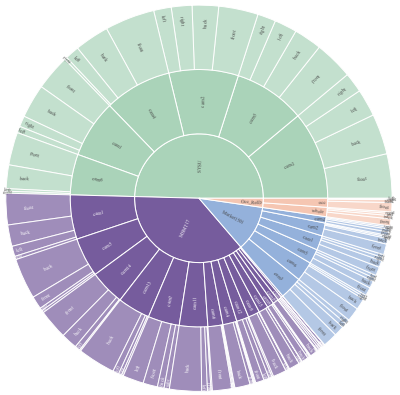

Figure 6. Distribution of cam and view in PersonSyn

Table 4. Quantitative performance comparison for pedestrian generation using the Market-1501 dataset.

| dataset | TransReID | | CLIP-ReID | |
|---|---|---|---|---|
| | mAP | rank-1 | mAP | rank-1 |
| Market | 87.3 | 94.3 | 89.8 | 95.5 |
| Market+DGGAN | 75.8 | 89.3 | 86.6 | 94.4 |
| Market+DPTN | 82.6 | 92.2 | 85.3 | 93.3 |
| Market+Pose2Id | 87.4 | 94.5 | 89.4 | 95.0 |
| Market+ours | **88.7** | **94.7** | **90.4** | **95.5** |

### 5.5. ReID Data Augmentation

To demonstrate the utility of OmniPerson for downstream tasks, we apply it to data augmentation for ReID. We enrich the Market-1501 training set by generating 4 standard-pose images for each identity. As shown in Table 6, training a standard ReID model on data augmented by OmniPerson yields stronger performance improvements compared to using data generated by other methods. It is important to note that our goal is to provide a unified generation tool, not a specialized augmentation method. We therefore used a simple augmentation strategy, which places a higher demand on the generation quality and identity consistency of the synthesized images. Even with this non-optimized approach, OmniPerson delivers stronger augmentation results than competing methods. For fairness, we note that other SoTA methods might also enhance ReID performance, rather than hinder it, if paired with their own tailored augmentation strategies.

### 5.6. Statistics and distribution of PersonSyn

We present a statistical analysis of the PersonSyn dataset. A subset of these distributions is visualized in Figure 6, **while a comprehensive breakdown is provided in the Supplementary Materials.**

## 6. Conclusion

In this paper, we propose OmniPerson, the first unified framework for pedestrian generation. OmniPerson integrates multi-reference image to maintain gelleray level identity consistency, while utilizing multi signals to enable visible/infrared pedestrian image/video generation. Furthermore, we introduce and will open-source PersonSyn, a novel multi-modal, multi-reference dataset for pedestrian generation. Experimental results demonstrate that OmniPerson achieves superior generation quality and identity consistency and effectively augments ReID datasets.

# OmniPerson: Unified Identity-Preserving Pedestrian Generation

## Supplementary Material

## A. PersonSyn Statistics and Details

We introduce **PersonSyn**, the pioneering large-scale dataset for **multi-reference, controllable** pedestrian generation. This section begins with a detailed description of the target image annotations and concludes with the dataset's composition.

### A.1. Annotation

For each target image, the annotations comprise three main components: conditional inputs, text annotations, and a reference image set. The detailed data annotation structure of PersonSyn is illustrated in Figure 7.

- Conditional Inputs: These include the skeleton image, SMPL-X parameters, rendered mesh images, and the background image.
- Text Annotations: These cover the orientation description, orientation vector, upper- and lower- garment, as well as attributes indicating whether the pedestrian is carrying a backpack, holding hand-carried items, or cycling.
- Reference Image Set: This includes the full collection of reference images (sorted by CLIP-ReID similarity to the target) and their categorizations based on camera identity, orientation, and pedestrian attributes (e.g., carrying a backpack).

### A.2. Composition

PersonSyn is constructed from several public large-scale ReID datasets, including Market1501, MSMT17, SYSU-MM01, Occluded-ReID, MARS, and HITSZ-VCM. These datasets cover a diverse range of tasks: Market1501 and MSMT17 represent traditional image-based ReID; SYSU-MM01 focuses on visible-infrared cross-modality ReID; Occluded-ReID targets occlusion challenges; MARS is designed for video-based ReID; and HITSZ-VCM is tailored for video-based visible-infrared ReID. Specifically, regarding Occluded-ReID, we employ whole-body images as the targets and occluded images as the references. For MARS, the video tracklets are assigned as target images, while Market1501 images are used to provide the reference information. As for HITSZ-VCM, we sample three frames from each video sequence—the start, end, and a random intermediate frame—to serve as the reference images. Detailed statistics on the camera and orientation distributions of the image-based ReID datasets are presented in Table 5, and the visualization results are shown in Figure 8.

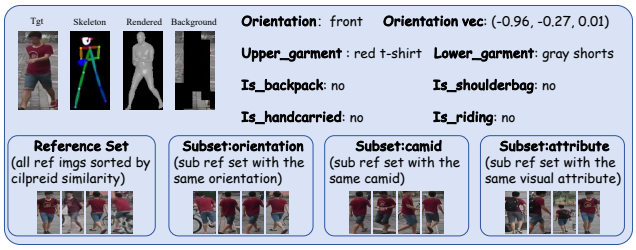


Figure 7. Annotation of PersonSyn

Table 5. Statistics of cameras and orientations in image-based ReID datasets.

| dataset | cam | all | front | back | left | right |
|---|---|---|---|---|---|---|
| Market1501 | 1 | 1568 | 625 | 492 | 207 | 244 |
| | 2 | 1399 | 775 | 401 | 119 | 104 |
| | 3 | 2451 | 1043 | 743 | 333 | 332 |
| | 4 | 688 | 244 | 204 | 128 | 112 |
| | 5 | 1580 | 598 | 528 | 222 | 232 |
| | 6 | 2130 | 994 | 815 | 155 | 166 |
| MSMT17 | 1 | 4783 | 2263 | 1540 | 740 | 240 |
| | 2 | 187 | 1 | 184 | 1 | 1 |
| | 3 | 448 | 127 | 151 | 136 | 34 |
| | 4 | 1523 | 260 | 1100 | 84 | 79 |
| | 5 | 4240 | 931 | 2935 | 166 | 208 |
| | 6 | 1623 | 1408 | 90 | 26 | 89 |
| | 7 | 3398 | 1040 | 478 | 1500 | 380 |
| | 8 | 726 | 69 | 481 | 155 | 21 |
| | 9 | 1253 | 245 | 960 | 22 | 26 |
| | 10 | 625 | 5 | 578 | 29 | 13 |
| | 11 | 3064 | 254 | 2305 | 412 | 93 |
| | 12 | 1335 | 607 | 229 | 389 | 110 |
| | 13 | 3513 | 228 | 2711 | 440 | 134 |
| | 14 | 3800 | 2156 | 1170 | 381 | 93 |
| | 15 | 1037 | 13 | 777 | 3 | 244 |
| SYSU-MM01 | 1 | 6240 | 2560 | 2425 | 476 | 779 |
| | 2 | 7488 | 2837 | 1901 | 1264 | 1486 |
| | 3 | 9502 | 3188 | 2910 | 1968 | 1436 |
| | 4 | 7264 | 3581 | 2480 | 1019 | 184 |
| | 5 | 7737 | 2785 | 2010 | 1627 | 1315 |
| | 6 | 4380 | 2357 | 1698 | 183 | 142 |
| dataset | occ | all | front | back | left | right |
| Occluded-ReID | occ | 969 | 600 | 214 | 45 | 50 |
| | whole | 1000 | 498 | 230 | 153 | 119 |

## B. Qualitative Results and Visualization

Figure 9 presents a qualitative comparison between OmniPerson and baseline methods. Supplementary generation results are provided in the subsequent figures. Specifically, **Figure 10 illustrates generation results of multiple iden-**


Figure 8. Statistical distribution of camera views and person orientations.

**tities in predefined poses.** In each group, the leftmost image serves as the real reference (ReID sample), followed by seven synthesized images rendered in fixed standard poses. These visualizations demonstrate that OmniPerson achieves precise control over pedestrian poses and Visible-Infrared modalities while faithfully preserving identity information. Furthermore, **Figure 11 showcases video generation results of multiple identities.** Similarly, the first image represents the real reference, followed by six frames sampled from a generated video sequence with continuous pose transitions. This highlights OmniPerson's capability to generate smooth video sequences characterized by natural motion. Notably, while maintaining robust identity consistency, the synthesized videos exhibit significantly larger variations in viewpoint and pose compared to existing real-world Video-ReID datasets. Finally, **Figure 12 displays generation results of random identities in random poses.** In each group, the first image serves as the real ReID reference, followed by a randomly generated target pose in the middle, and the final synthesized image. These generated results encompass visible image generation, visible-to-infrared generation, and infrared-to-infrared generation, demonstrating the high diversity and coverage of our generative model.

As a multi-reference generation framework, the primary advantage of OmniPerson over the baseline is its ability to maintain gallery consistency via multiple references, rather than merely preserving consistency with a single reference image. Ablation studies on the number of reference images are presented in Section D.2


Figure 9. Generation quality comparison: Baseline vs. ours.

## C. Algorithmic and Implementation Details

### C.1. Pedestrian orientation vector calculation

We extract and compute orientation unit vector from smpl parameters. The rotation matrix is given by the Rodrigues formula:

$$\mathbf{R} = \mathbf{I} + \sin\theta\,[\mathbf{n}]_\times + (1-\cos\theta)\,[\mathbf{n}]_\times^2 \tag{2}$$

where $\theta$ is smpl root pose and matrix $[\mathbf{n}]_\times$ defined as:

$$[\mathbf{n}]_\times = \begin{bmatrix} 0 & -n_z & n_y \\ n_z & 0 & -n_x \\ -n_y & n_x & 0 \end{bmatrix} \tag{3}$$

The final pedestrian orientation unit vector is obtained as:

$$\mathbf{v}_{\text{ori}} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} r_{13} \\ r_{23} \\ r_{33} \end{bmatrix} \tag{4}$$

**Algorithm 1** SMPL Parameters Interpolation

**Input**: SMPL-X parameters $\mathbf{para}_1$, $\mathbf{para}_2$; interpolation coefficient $\alpha \in [0, 1]$

**Output**:paras $\mathbf{para}_{\text{interp}}$

Initialize empty parameter dictionary $\mathbf{para}_{\text{interp}}$
**for** each key $k$ in $\mathbf{para}_1$ **do**
  **if** $k$ is 'global_orient' **then**
    {Handle rotation with quaternion SLERP}
    Convert $\mathbf{para}_1[k]$, $\mathbf{para}_2[k]$ to rotation vectors $\mathbf{r}_1, \mathbf{r}_2$
    Convert $\mathbf{r}_1, \mathbf{r}_2$ to quaternions $\mathbf{q}_1, \mathbf{q}_2$
    **if** $\mathbf{q}_1 \cdot \mathbf{q}_2 < 0$ **then**
      $\mathbf{q}_2 \leftarrow -\mathbf{q}_2$ {Ensure shortest path}
    **end if**
    Compute spherical interpolation: $\mathbf{q}_{\text{interp}} = \frac{\sin[(1-\alpha)\theta]}{\sin\theta}\mathbf{q}_1 + \frac{\sin[\alpha\theta]}{\sin\theta}\mathbf{q}_2$
    where $\theta = \arccos(\mathbf{q}_1 \cdot \mathbf{q}_2)$
    Convert $\mathbf{q}_{\text{interp}}$ back to rotation vector
    Store result in $\mathbf{para}_{\text{interp}}[k]$
  **else**
    {Linear interpolation for other parameters}
    $\mathbf{para}_{\text{interp}}[k] \leftarrow (1-\alpha)\mathbf{para}_1[k] + \alpha\mathbf{para}_2[k]$
  **end if**
**end for**
**return** $\mathbf{para}_{\text{interp}}$

### C.2. Image-guided sequence interpolation method

We generate gradually varying pedestrian parameters through SMPL parameter interpolation and render them to obtain pedestrian video guidance sequences. The interpolation pseudocode is provided in Algorithm 1

### C.3. Reference Image Selection Strategy.

During training, we select reference images based on the following criteria:

- **Same-camid Reference**: We explicitly select at least one image from the same camera as the target. Beyond serving as a visual reference, this image is utilized to extract ReID features for guiding generation and to extract the background, which serves as a background prompt.
- **Dynamic Selection**: For the remaining references, we adopt a stage-aware strategy. In the early training stage, we select images with the highest CLIP-ReID similarity to the target from each orientation group. In the later stage, we impose a viewpoint constraint (e.g., requiring a difference of at least $60^\circ$ from the target view) to filter the candidate pool. The candidates are then sorted by ReID similarity, and multiple references are selected based on Bernoulli probabilities.

Table 6. Effectiveness of data augmentation on SYSU-MM01

| Method | dataset | R-1↑ | R-10↑ | R-20↑ | mAP↑ | mInp↑ |
|---|---|---|---|---|---|---|
| AGW[48] | SYSU-MM01 | 47.5 | 84.4 | 92.1 | 47.7 | 35.3 |
| HAT[47] | SYSU-MM01 | 55.3 | 92.1 | 97.4 | 53.9 | - |
| LBA[31] | SYSU-MM01 | 55.4 | 91.1 | - | 54.1 | - |
| NFS[6] | SYSU-MM01 | 56.9 | 91.3 | 96.5 | 55.5 | - |
| MSO[13] | SYSU-MM01 | 58.7 | 92.1 | 97.2 | 56.4 | 42.0 |
| CM-NAS[12] | SYSU-MM01 | 62.0 | 92.9 | 97.3 | 60.0 | - |
| MID[19] | SYSU-MM01 | 60.3 | 92.9 | - | 95.4 | - |
| SPOT[5] | SYSU-MM01 | 65.3 | 92.7 | 97.0 | 62.3 | 48.9 |
| FMCNet[56] | SYSU-MM01 | 66.3 | - | - | 62.5 | - |
| PMT[27] | SYSU-MM01 | 67.5 | 95.4 | 98.6 | 64.9 | 51.9 |
| | SYSU+ours | **68.3** | **96.0** | **99.0** | **66.7** | **54.1** |

Table 7. Ablation Study on the Maximum Number of Ref-Img

| Max-Ref-Img | LPIPS↓ | SSIM↑ | PSNR↑ |
|---|---|---|---|
| 1 | 0.270 | 0.284 | 13.833 |
| 2 | 0.235 | 0.359 | 15.625 |
| 3 | 0.216 | 0.437 | 16.922 |
| 4 | 0.210 | 0.455 | 16.912 |

### C.4. Data Augmentation

For the ReID data used to train the diffusion model, we implement the following augmentation techniques:

- **Random cropping** is applied to both reference and target images before other transformations.
- **Random occlusion** is introduced specifically to the reference images.
- **Background masking** is performed with a high probability (i.e., randomly masking out the background image).

## D. Additional Quantitative Evaluations

### D.1. Visual-Infrared ReID Data Augentation

As illustrated in Table 6, we generated a video sequence with continuously changing actions for each pedestrian under every camera view in the SYSU-MM01 dataset and incorporated them into the training set. We then trained the ReID model and evaluated its performance under the 'all-search' mode. Experimental results demonstrate that the visible-infrared data generated by OmniPerson effectively enhances VI-ReID performance.

### D.2. Ablation for Multi Reference Images

Table 7 presents the quantitative metrics for ReID generation across varying maximum reference limits. As shown in the table, a higher number of reference images contributes significantly to better generation quality. We attribute this improvement to the gallery consistency guaranteed by our multi-reference mechanism. By avoiding the bias inherent in aligning with a single reference image, the model achieves higher fidelity. Furthermore, this mechanism allows for the inclusion of historically generated images into the reference set, ensuring consistency across the entire generation sequence.

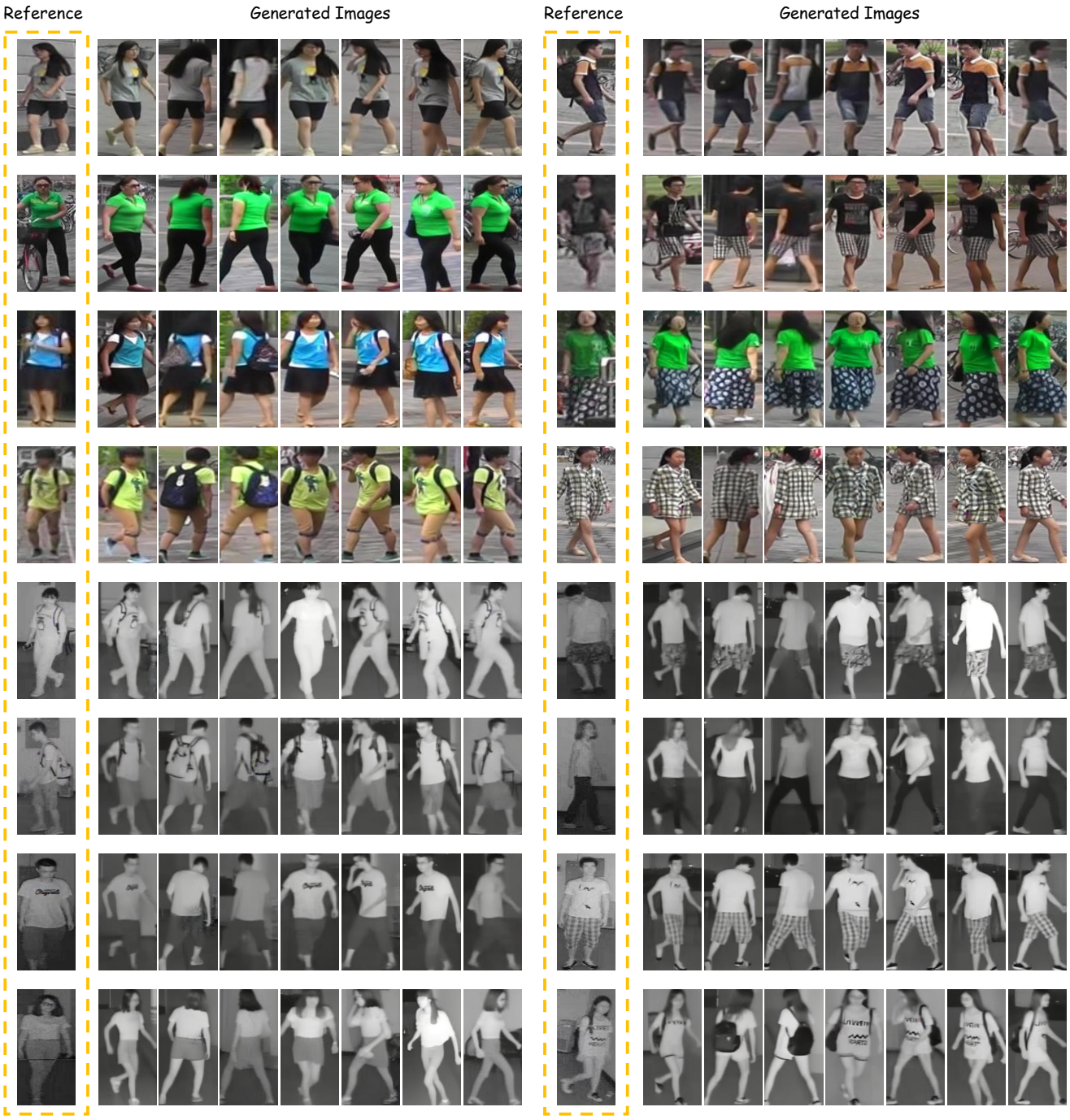


Figure 10. Generation results of multiple identities in predefined poses


Figure 11. Video generation results of multiple identities

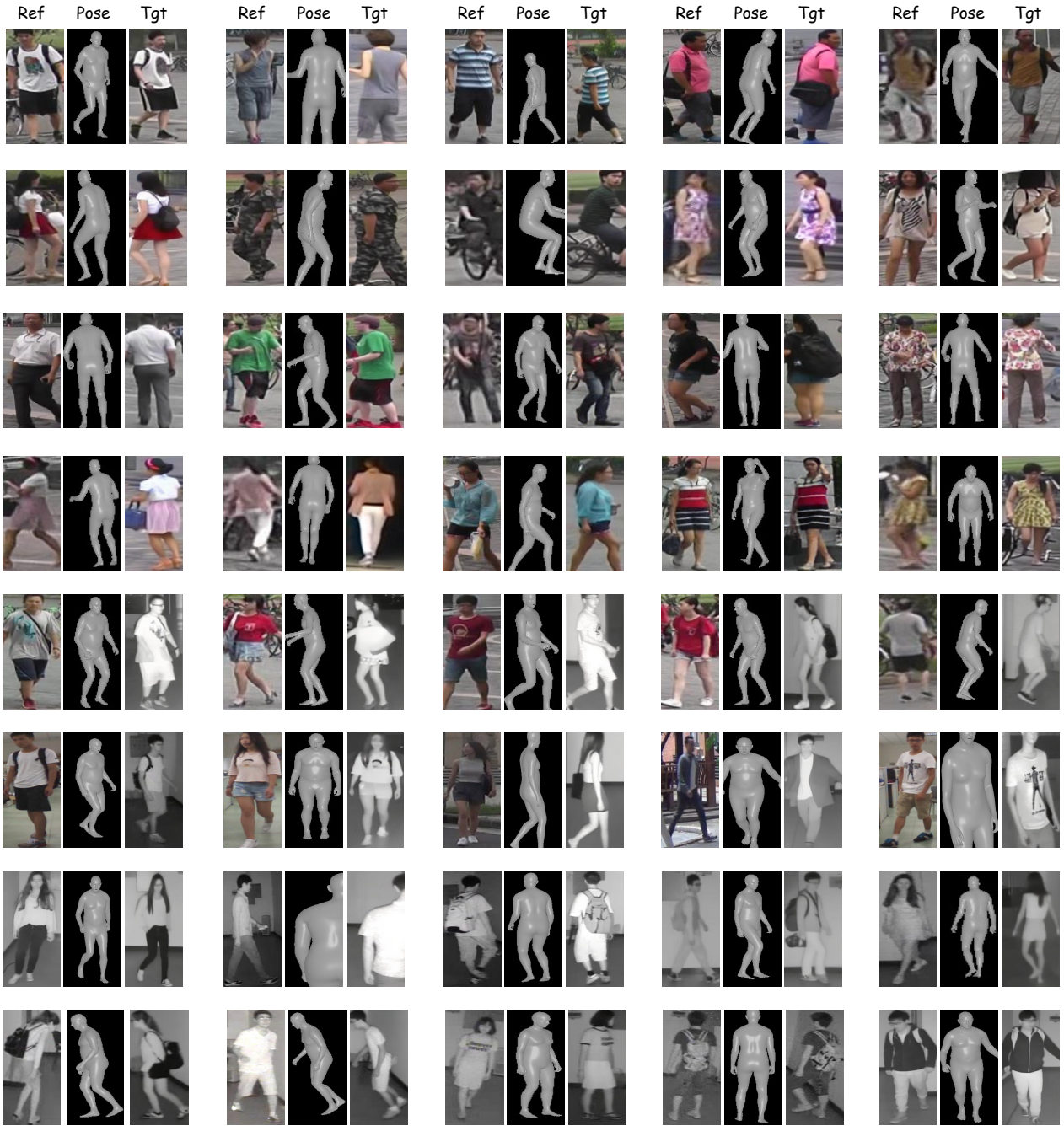


Figure 12. Generation results of random identities in random poses

## References


[1] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 2, 3

[2] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023. 3

[3] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023. 3

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 3

[5] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 31:2352–2364, 2022. 3

[6] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 587–597, 2021. 3

[7] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, page 6, 2018. 3

[8] Wenbo Dai, Lijing Lu, and Zhihang Li. Diffusion-based synthetic data generation for visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11185–11193, 2025. 2

[9] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. Idm: An intermediate domain module for domain adaptive person re-id. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11864–11874, 2021. 2

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3

[11] Yu Dong, Yihao Liu, He Zhang, Shifeng Chen, and Yu Qiao. Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10729–10736, 2020. 7

[12] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11823–11832, 2021. 3

[13] Yajun Gao, Tengfei Liang, Yi Jin, Xiaoyan Gu, Wu Liu, Yidong Li, and Congyan Lang. Mso: Multi-feature space joint optimization network for rgb-infrared person re-identification. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5257–5265, 2021. 3

[14] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *Advances in neural information processing systems*, 31, 2018. 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[16] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15013–15022, 2021. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[18] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8153–8163, 2024. 5, 6, 7

[19] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1034–1042, 2022. 3

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 3

[21] Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary. Person re-identification by attributes. In *Bmvc*, page 8, 2012. 3

[22] He Li, Mang Ye, and Bo Du. Weperson: Learning a generalized re-identification model from all-weather virtual data. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3115–3123, 2021. 3

[23] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1405–1413, 2023. 2

[24] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3693–3702, 2019. 7

[25] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 456–474. Springer, 2020. 3

[26] Zanwu Liu, Chao Yuan, Bo Li, Xiaowei Zhang, and Guanglin Niu. Looking alike from far to near: Enhancing cross-resolution re-identification via feature vector panning. *arXiv preprint arXiv:2510.00936*, 2025. 3

[27] Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the*

*AAAI conference on artificial intelligence*, pages 1835–1843, 2023. 3

[28] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10806–10815, 2021. 7

[29] Ke Niu, Haiyang Yu, Xuelin Qian, Teng Fu, Bin Li, and Xiangyang Xue. Synthesizing efficient data with diffusion models for person re-identification pre-training. *Machine Learning*, 114(3):1–25, 2025. 2, 3

[30] Zhiqi Pang, Jifeng Guo, Wenbo Sun, Yanbang Xiao, and Ming Yu. Cross-domain person re-identification by hybrid supervised and unsupervised learning. *Applied Intelligence*, 52(3):2987–3001, 2022. 3

[31] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12046–12055, 2021. 3

[32] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 3

[33] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–667, 2018. 2, 3

[34] Weichao Qiu and Alan Yuille. Unrealcv: Connecting computer vision to unreal engine. In *European conference on computer vision*, pages 909–916. Springer, 2016. 3

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[36] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7690–7699, 2020. 7

[37] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 2

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, and et al. Whye Teoh. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2205.11487. 2

[41] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018. 7

[42] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 3800–3808, 2017. 3

[43] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *European Conference on Computer Vision*, pages 717–734. Springer, 2020. 2, 3, 7

[44] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[45] Yanan Wang, Xuezhi Liang, and Shengcai Liao. Cloning outfits from real-world images to 3d characters for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4900–4909, 2022. 3

[46] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017. 3

[47] Mang Ye, Jianbing Shen, and Ling Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, 16:728–739, 2020. 3

[48] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 3

[49] Chao Yuan, Zanwu Liu, Guiwei Zhang, Haoxuan Xu, Yujian Zhao, Guanglin Niu, and Bo Li. Modality-transition representation learning for visible-infrared person re-identification. *arXiv preprint arXiv:2511.02685*, 2025. 3

[50] Chao Yuan, Guiwei Zhang, Changxiao Ma, Tianyi Zhang, and Guanglin Niu. From poses to identity: Training-free person re-identification via feature centralization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24409–24418, 2025. 2, 3, 7

[51] Chao Yuan, Tianyi Zhang, and Guanglin Niu. Neighbor-based feature and index enhancement for person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5762–5769, 2025. 3

[52] Guiwei Zhang, Yongfei Zhang, and Zichang Tan. Protohpe: Prototype-guided high-frequency patch enhancement

for visible-infrared person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 944–954, 2023.

[53] Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14133–14142, 2023. 3

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 5

[55] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7713–7722, 2022. 7

[56] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7349–7358, 2022. 3

[57] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Unrealperson: An adaptive pipeline towards costless person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11506–11515, 2021. 2, 3

[58] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 3

[59] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, 28(3):1176–1190, 2018. 3

[60] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 2

[61] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2347–2356, 2019. 3, 7