Pose-dIVE: Pose-Diversified Augmentation for Person Re-Identification

Inès Hyeonsu Kim[1*] Woojeong Jin[1*] Soowon Son[1] Junyoung Seo[1] Seokju Cho[1]
JeongYeol Baek[2] Byeongwon Lee[2] JoungBin Lee[1] Seungryong Kim[1†]
KAIST AI[1] SK Telecom[2]

Abstract

Person re-identification (Re-ID) often faces challenges due to variations in human poses and camera viewpoints, which significantly affect the appearance of individuals across images. Existing datasets frequently lack diversity and scalability in these aspects, hindering the generalization of Re-ID models to new camera systems or environments. To overcome this, we propose ***Pose-dIVE****, a novel data augmentation approach that incorporates sparse and underrepresented human pose and camera viewpoint examples into the training data, addressing the limited diversity in the original training data distribution. Our objective is to augment the training dataset to enable existing Re-ID models to learn features unbiased by human pose and camera viewpoint variations. By conditioning the diffusion model on both the human pose and camera viewpoint through the SMPL model, our framework generates augmented training data with diverse human poses and camera viewpoints. Experimental results demonstrate the effectiveness of our method in addressing human pose bias and enhancing the generalizability of Re-ID models compared to other data augmentation-based Re-ID approaches. Our project page is available at:* https://cvlab-kaist.github.io/Pose-dIVE.

1. Introduction

Person re-identification (Re-ID) is widely utilized in modern surveillance systems to track and recognize individuals across multiple camera networks [7, 44, 56, 57]. Despite significant advancements in Re-ID methodologies [5, 22, 46, 62], there remains a notable gap between performance under controlled training conditions and effectiveness in real-world scenarios.

Two particularly challenging factors limiting generalization are changes in human pose [8, 38, 52] and variations in camera viewpoint [1, 18]. Although an individual's identity remains constant, differences in pose or camera angle can significantly alter their visual appearance. Thus, robust Re-ID models must capture identity-defining features despite these spatial variations. However, current Re-ID datasets often lack sufficient diversity in poses and viewpoints. Typically, these datasets feature only limited walking or standing poses [55] and employ just two or three camera viewpoints per identity [23, 45, 54]. Such restricted diversity produces unimodal samples that hinder models from learning robust identity representations. Furthermore, testing sets that lack diverse scenarios may not accurately reflect real-world complexities.

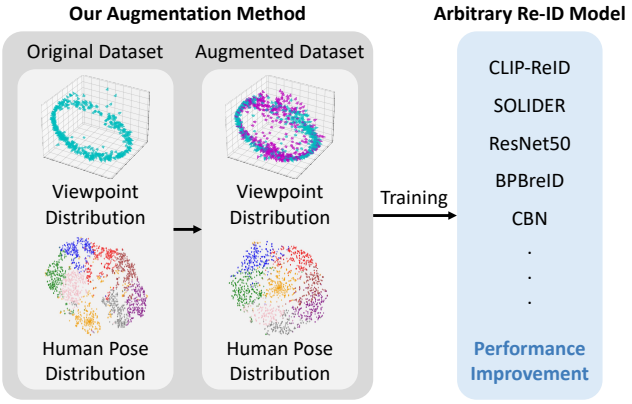


Figure 1. **Pose-dIVE** diversifies the viewpoint and human pose of Re-ID datasets to help generalize and improve the performance of arbitrary Re-ID models. On a curated real-world test dataset, the model [10] trained with the Pose-dIVE augmented dataset achieves a **+13.6 mAP** and **+11.0 R1** improvement (refer Table 6).

Collecting and annotating richer and more varied datasets could mitigate these issues. Nevertheless, privacy concerns and the high cost associated with large-scale, multi-view camera installations [10, 27, 53] further complicate the collection of richer, more varied datasets. As a result, pose invariant feature learning [11, 18, 19, 25] and data augmentation [3, 4, 30, 33, 58, 61] have become critical approaches in addressing this limitation. Existing augmentation techniques primarily exploit the limited range of poses [11, 26] and viewpoints [3, 4] already present in cur-

*Equal contributions.
†Corresponding author.

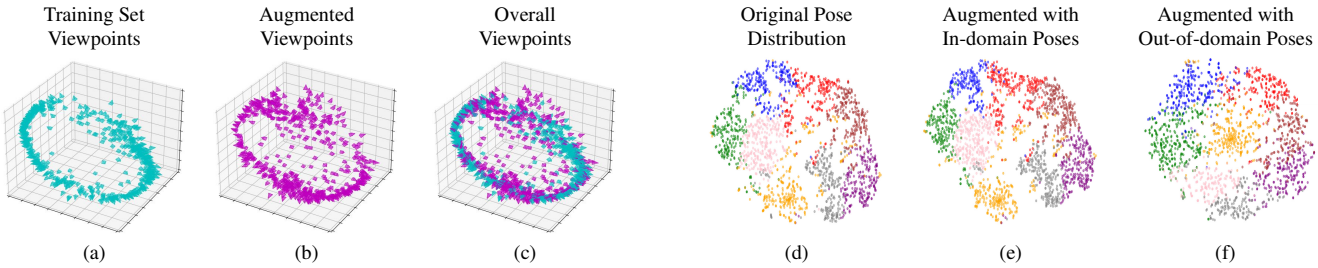


Figure 2. **Visualization of the effect of viewpoint and human pose augmentation.** We compare visualizations of camera viewpoint and human pose distributions for the Market-1501 [54]. The left figures (i) display the camera viewpoint distribution derived from SMPL, while the right figures (ii) illustrate the pose distribution. In (i), from left to right, we show the viewpoint distributions of the training dataset, the augmented dataset, and the combination of both. Similarly, in (ii), from left to right, we present t-SNE [41] visualizations of the human pose distributions, showing poses from the training dataset, followed by augmented poses sourced from outside the dataset. These visualizations demonstrate that Pose-dIVE successfully diversifies both viewpoint and human pose distributions.

rent datasets. Additionally, these methods usually restrict viewpoint augmentation to horizontal rotations and neglect elevation changes [3, 4]. Moreover, pose and viewpoint have traditionally been treated as separate factors, despite their combined influence on human appearance.

To address these issues, our work proposes an integrated augmentation strategy that jointly considers pose and viewpoint variations while introducing greater variability in both factors, as illustrated in Figure 2. Instead of relying solely on poses from existing Re-ID datasets, we incorporate dynamic poses sourced externally. We simultaneously adjust both azimuth and elevation angles to achieve a broader and more realistic range of camera viewpoints. This comprehensive augmentation strategy enables Re-ID models to better recognize stable identity features across varied appearances. Our experiments validate this approach by demonstrating improved performance even when existing models are trained with these augmented samples.

Our approach utilizes recent large-scale diffusion models [37], leveraging their ability to encode extensive prior knowledge about human appearances under diverse conditions. By conditioning these models within a dual-branch architecture [15] that simultaneously preserves reference identity and integrates SMPL-derived [28] pose and viewpoint guidance, we generate high-fidelity training samples. This specifically targets distributional gaps found in conventional Re-ID datasets.

To rigorously evaluate model robustness, we test systems trained with our augmented dataset on both real-world data and a customized evaluation set containing poses and viewpoints absent from the training data. This methodology effectively measures how well models generalize to unseen conditions and practical scenarios. As shown in Figure 1, our experiments confirm that training on datasets augmented by Pose-dIVE significantly enhances Re-ID performance on standard benchmarks and in more challenging real-world conditions.

2. Related Work

Data augmentation in person re-identification. Person re-identification (Re-ID) has advanced significantly with deep learning techniques, enabling robust matching of individuals across non-overlapping camera views [5, 22, 39]. However, a persistent challenge in Re-ID is the scarcity of diverse training datasets, which limits model generalization. This issue has spurred exploration into various strategies, including data augmentation, to enhance the robustness and scalability of Re-ID systems.

Early augmentation methods adopted straightforward image transformations such as random resizing, cropping, and horizontal flipping [29], as well as techniques to simulate occlusions [16]. These approaches enrich the training data, providing benefits to methods like distance metric learning [21, 24], which rely on diverse samples to better embed features of the same identity closely together while separating those of different identities. While effective in increasing data variety, these techniques primarily leverage variations already present in the dataset, leaving gaps in addressing underrepresented aspects such as human pose and camera viewpoint diversity.

Generative data augmentation in Re-ID. More advanced methods have employed Generative Adversarial Networks (GANs) [12] to synthesize training images. LSRO [58] utilizes DCGANs [34] to generate unlabeled samples, enhancing semi-supervised Re-ID through label smoothing regularization. Similarly, pose-transferrable GANs [26] augment datasets by transferring poses, though they rely on poses extracted from existing Re-ID data. Other works, such as [59], focus on generating cross-ID images to support joint discriminative and generative learning. These GAN-based approaches have demonstrated performance improvements by expanding the training data. However, their emphasis often lies in utilizing variations already present within the dataset or addressing objectives such as labeling or identity synthesis. In contrast, our work focuses

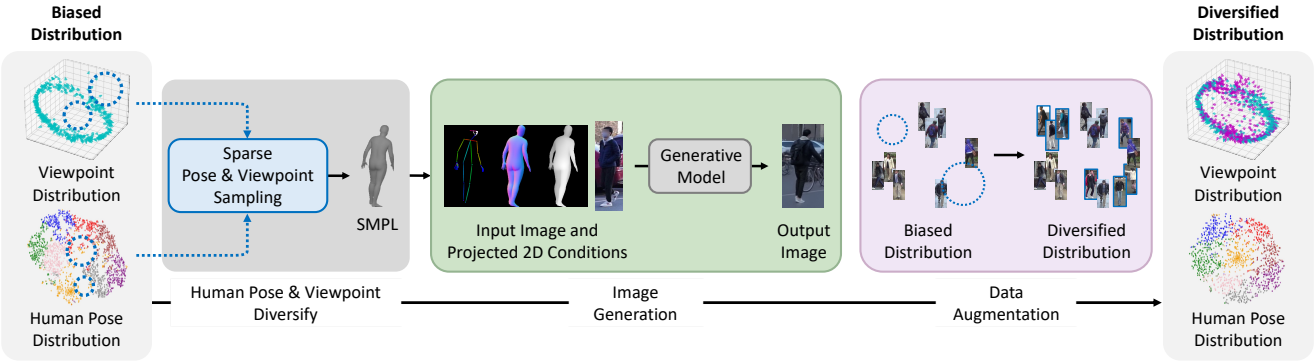


Figure 3. **Pose-dIVE framework.** Upon observing the highly biased viewpoint and human pose distributions in the existing training dataset, we augment the dataset by manipulating SMPL body shapes and feeding the rendered shapes into a generative model to fill in sparsely distributed poses and viewpoints. With this augmented dataset, we can train a Re-ID model that is robust to viewpoint and human pose biases.

on diversifying the training dataset by uniformly sampling viewpoints and human poses from external datasets.

3D mesh guidance in Re-ID. Several studies have utilized 3D human mesh representations, such as SMPL [28], to model human body shape in Re-ID tasks. For instance, GCL [3] employs a 3D mesh-based view generator that rotates the mesh horizontally to create new viewpoints while preserving the given human pose. This preservation of the original pose contrasts with our objective, as we aim to diversify both poses and viewpoints in the training data. Other efforts, such as 3DInvarReID [25], focus on long-term Re-ID and 3D body shape reconstruction, which are distinct tasks from our goal of enhancing training data diversity. Similarly, Chen et al. [4] uses a 3D mesh to guide GANs for unsupervised Re-ID, but it relies on poses and viewpoints already available in the Re-ID dataset. Our approach, however, integrates external pose and viewpoint data, setting it apart from these in-domain strategies.

3. Method

3.1. Overview

In this paper, we propose a data augmentation strategy designed to address the limitations of existing Re-ID datasets, particularly the restricted range of camera viewpoints and human poses, which is depicted in Figure 3. In this section, we first explain the augmentation process, detailing how camera viewpoints and human poses are represented during augmentation. Next, we describe how these conditions, along with identity information, are incorporated into the diffusion model, which is designed to accommodate these conditions. The data generated using our approach can be applied to any Re-ID model, enhancing its generalizability.

3.2. Human Pose and Camera Viewpoint Condition with SMPL

A key component of our augmentation strategy is the ability to generate diverse human poses and camera viewpoints in a controlled manner. To accomplish this, we provide specific conditions to the generative model to guide the augmentation process.

Previous works [11, 32, 40] have primarily addressed pose control by providing a human pose skeleton. However, relying solely on the human skeleton has limitations due to missing information: when projected onto 2D images, the human skeleton lacks depth information, posing ambiguity for the model when inferring viewpoint information from the skeleton. For example, if the camera is above a person, the skeleton would appear compressed vertically. Without depth information, the model cannot distinguish whether the camera is positioned above the person or the person is simply short.

In this regard, in addition to the human pose skeleton, we utilize SMPL [28], a human body model used for realistic human rendering. SMPL can model intricate human shapes, including complex body articulations in 3D space. From this human model, we can easily extract 2D representations of the 3D human by rendering the model, such as depth maps which implicitly contain camera viewpoint information.

Let us define the SMPL model as a function $\mathcal{M}$ that generates a 3D mesh based on shape parameters $\beta \in \mathbb{R}^{10}$ and pose parameters $\theta \in \mathbb{R}^{72}$:

$$\{\mathcal{V}, \mathcal{F}\} \leftarrow \mathcal{M}(\beta, \theta), \tag{1}$$

where $\mathcal{V} \in \mathbb{R}^{N\times 3}$ represents the set of N vertices in 3D space and $\mathcal{F} \in \mathbb{N}^{F\times 3}$ represents the set of F triangular faces, each defined by three vertex indices from $\mathcal{V}$.

To extract 2D representations from this 3D model, we define a rendering function $\mathcal{R}$ that projects the 3D mesh onto a 2D plane given camera parameters $\phi = (R, t, K)$, where $R \in \mathbb{R}^{3\times 3}$ is the rotation matrix, $t \in \mathbb{R}^3$ is the translation vector, and $K \in \mathbb{R}^{3\times 3}$ is the intrinsic camera matrix:

$$\{I_d, I_n, I_s\} \leftarrow \mathcal{R}(\{\mathcal{V}, \mathcal{F}\}, \phi), \tag{2}$$

where $I_d \in \mathbb{R}^{H\times W\times 1}$ is the depth map, $I_n \in \mathbb{R}^{H\times W\times 3}$ is the surface normal map, and $I_s \in \mathbb{R}^{H\times W\times 3}$ is the skeleton representation with J joints.

To enrich the camera viewpoint information, we incorporate depth maps. Additionally, surface normals from SMPL are used to capture detailed human surface characteristics, enhancing the precision of the generated augmentations. As a result, depth maps, surface normals, and human skeletons from SMPL, along with a reference image to control identity, are fed into the generative model as guidance.

3.3. Pose and Viewpoint Diversification

To mitigate the biased camera viewpoint and human pose in training dataset, we augment images with *uniformly distributed camera viewpoint* and *human poses sourced from outside the training dataset*. For camera viewpoints, we augment the images adjusting two factors: elevation and azimuth.

We sample the elevation angle α from a uniform distribution with the hyperparameters $\alpha_{\min}$ and $\alpha_{\max}$, denoted as $\alpha \sim \mathcal{U}(\alpha_{\min}, \alpha_{\max})$, assuming that the camera is not positioned below the ground and that a person becomes indistinguishable when the camera is positioned above a certain degree. Similarly, for the azimuth angle γ, we uniformly sample within the bounds $\gamma_{\min}$ and $\gamma_{\max}$, represented as $\gamma \sim \mathcal{U}(\gamma_{\min}, \gamma_{\max})$. In addition, the camera is always directed towards the center of the human mesh. This approach allows for the capture of a person from any direction. With these considerations, we found that the distribution of camera viewpoint is significantly unbiased, as depicted in Figure 2.

For human poses, we source them from outside the training dataset for diversification. As Re-ID datasets often contain similar human poses, solely training on these datasets can lead to overfitting and limited generalization capabilities. Let $\mathcal{P}_{\text{ext}}$ represent the set of human poses extracted from external sources (e.g., dance videos [2]). For each augmentation, we randomly sample a pose $\theta \sim \mathcal{P}_{\text{ext}}$ from this external distribution uniformly. By incorporating a wide range of external poses, our approach improves the model's ability to handle unseen poses that are not present in the training dataset. Using these diversified viewpoint and pose distributions, we render SMPL models into 2D representations according to:

$$\{I_d, I_n, I_s\} = \mathcal{R}(\mathcal{M}(\beta, \theta), \phi), \tag{3}$$

where $\phi = (R(\alpha, \gamma), t, K)$, $R(\alpha, \gamma)$ represents rotation matrix derived from azimuth and elevation, $\alpha \sim \mathcal{U}(\alpha_{min}, \alpha_{max})$, and $\gamma \sim \mathcal{U}(\gamma_{min}, \gamma_{max})$, respectively. These rendered 2D representations act as input conditions for the generative model, providing guidance for the data generation process. With the augmented training dataset from generative model, we can train arbitrary Re-ID models with robustness to camera viewpoint and human pose variations.

3.4. Pose-Diversified Augmentation

We generate training images with diverse camera viewpoints and human poses to reduce bias in the distribution of the training data. However, when training the generative model on a human Re-ID dataset without careful consideration, it may produce poor results for camera viewpoints or human poses that are rarely present in the training dataset, as the quality of the generated data is limited by the capabilities of the generative model. If the generative model is unable to handle out-of-distribution human poses not seen during training, it will produce degraded training data.

In this work, we address this problem by leveraging the extensive knowledge in pre-trained Stable Diffusion (SD) [37]. Specifically, we fine-tune SD to accommodate rendered pose conditions, adapting the framework proposed by Hu et al. [15]. This approach effectively preserves the identity of the input image while taking advantage of Stable Diffusion's comprehensive pre-trained knowledge.

Let x_0 be the target image we want to generate, and x_T be pure Gaussian noise. The forward process gradually adds noise to the image according to:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{4}$$

where β_t is the noise schedule at timestep t.

The approach clones the pre-trained diffusion model into two branches of U-Nets. One branch, a reference U-Net ϵ_θ^{ref}, receives an image of a person whose identity is to be generated, while the other is a denoising U-Net ϵ_θ that gradually removes Gaussian noise according to:

$$p_\theta(x_{t-1}|x_t, c_{\text{pose}}) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c_{\text{pose}}), \Sigma_\theta(x_t, t)), \tag{5}$$

where c_{pose} represents the conditioning information, and μ_θ is derived from the predicted noise ϵ_θ.

A reference U-Net provides the identity information to the denoising U-Net through an attention mechanism. Identity information is shared with the denoising U-Net within self-attention [42] in each block, allowing the two parallel branches to benefit from the comprehensive pre-trained knowledge of Stable Diffusion. Formally, for each attention layer in the network:

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

where d_k denotes the scaling factor, the query Q comes from the denoising branch, while the key K and value V are derived from both the reference and denoising branch to provide identity guidance while denoising.

In addition, we concatenate the depth, surface normals, and skeleton, and feed them into the pose guider network G, which consists of stacks of convolutional layers:

$$c_{\text{pose}} = G([I_d, I_n, I_s]), \tag{7}$$

where $[\cdot]$ denotes concatenation. The encoded condition is then added to the projected input of the denoising diffusion model, following [15].

The diffusion model can generate an image that retains the identity from the reference image while allowing control over its viewpoint and human pose. Augmenting the training dataset with the diffusion model conditioned on the pose distribution outlined in Sec. 3.3 helps reduce bias when training a Re-ID model. For a detailed description of the architecture, please refer to the supplementary materials.

4. Experiments

4.1. Implementation Details

We use CLIP-reID [22] and SOLIDER [5] as the baselines to validate the effectiveness of the Pose-dIVE augmented dataset. The entire training process is divided into three parts: training the generative model, generating images for augmentation, and training the baseline Re-ID models using the augmented dataset.

Step 1: Training of generative model. The training of our generative model involves two stages. First, it learns a general human representation using a fashion video dataset [49], followed by fine-tuning on person Re-ID datasets. Throughout both stages, the weights of the autoencoders [48] and the CLIP [35] image encoder are kept frozen, focusing on the learning of reference U-Net, denoising U-Net, and pose guider. Initially, the reference and denoising U-Nets are initialized from the Stable Diffusion [37] model and fine-tuned with the fashion video dataset. We employ Mean Squared Error (MSE) loss, optimized using Adam [20] with a learning rate of 1e-5 and weight decay of 0.01. The first stage takes approximately 15 hours on a single NVIDIA RTX A6000 GPU with a batch size of 2. In the second stage, the model is further fine-tuned using Re-ID datasets consisting of pedestrian images from CCTV cameras, with the images resized to 192×384 and the batch size increased to 4.

Step 2: Augmentation with generative model. Our approach leverages data augmentation via a generative model to enrich the training data for the Re-ID model. We begin by rendering SMPLs extracted from the Everybody Dance Now dataset [2], an external dataset not used during Re-ID model training. From these SMPLs, we generate skeleton, depth, and normal maps using various camera viewpoints and a wide range of human poses, creating a "target condition gallery." For each person identity (PID), we randomly select one instance and pair it with a randomly chosen target condition from this gallery. This pair is then fed into the generative model to produce a new image. We repeat this process iteratively, ensuring the same number of iterations for each PID. These generated images are then combined with the original training datasets for training. The camera viewpoints of generated images are controlled by the azimuth γ and elevation angles α, where the angles are sampled within a limited range $\alpha_{min} = 0°, \alpha_{max} = 30°, \gamma_{min} = 0°$, and $\gamma_{max} = 360°$.

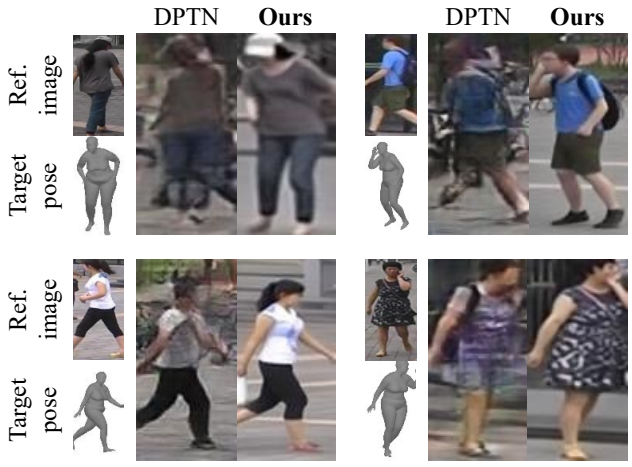


Figure 4. **Qualitative comparison.** We compare our generated output with DPTN [51], showing that Pose-dIVE can generate more realistic images while better preserving identity (e.g., cap and backpack) and accurately following the target pose.

Step 3: Training baseline Re-ID models. We trained two baseline Re-ID models, CLIP-reID and SOLIDER, using both the original and augmented Re-ID datasets. For training the baselines, we followed the procedure outlined in their respective papers.

4.2. Evaluation Protocol

We evaluate the performance of the models using four publicly available person Re-ID datasets, *i.e.*, MSMT17 [45], Market-1501 [54], CUHK03 (L), and CUHK03 (D) [23]. L and D stand for Labeled and Detected, respectively. For evaluation metrics, we use two standard Re-ID metrics: cumulative matching characteristics at Rank-1 (R1) and mean average precision (mAP).

4.3. Quantitative Comparisons

The quantitative results, presented in Table 1, demonstrate the effectiveness of our approach. To emphasize its applicability to arbitrary Re-ID models, we conducted experiments on two state-of-the-art models [5, 22]. Both models exhibited significant performance boosts when trained on a dataset augmented with our approach. This validates the broad applicability of our augmentation method to a variety of Re-ID models. Furthermore, we observed consistent per-

Methods	MSMT17		Market1501		CUHK03 (D)		CUHK03 (L)	
	mAP ↑	R1 ↑	mAP ↑	R1 ↑	mAP ↑	R1 ↑	mAP ↑	R1 ↑
TransReID [14]	69.4	86.2	89.5	95.2	-	-	-	-
AAFormer [63]	65.6	84.4	88.0	95.4	77.2	78.1	79.0	80.3
AGW [47]	49.3	68.3	87.8	95.1	-	-	62.0	63.6
FlipReID [31]	68.0	85.6	89.6	95.5	-	-	-	-
CAL [36]	64.0	84.2	89.5	95.5	-	-	-	-
PFD [43]	64.4	83.8	89.7	95.5	-	-	-	-
SAN [17]	55.7	79.2	88.0	96.1	74.6	79.4	76.4	80.1
LDS [50]	67.2	86.5	90.4	95.8	-	-	-	-
MPN [9]	62.7	83.5	90.1	96.4	79.1	83.4	81.1	85.0
MSINet [13]	59.6	81.0	89.6	95.3	-	-	-	-
SCSN [6]	58.5	83.8	88.5	95.7	81.0	84.7	84.0	86.8
Baseline (CLIP-reID [22])	68.0	85.8	89.6	95.5	93.7	95.5	95.5	96.6
+ Pose-dIVE	**71.0**	**87.5**	90.3	95.6	95.5	**97.4**	97.2	97.8
Baseline (SOLIDER [5])	67.4	85.9	91.6	96.1	95.6	96.7	97.4	**98.5**
+ Pose-dIVE	68.3	85.9	**92.3**	**96.6**	**96.2**	97.2	**97.6**	**98.5**

Table 1. **Quantitative comparison on standard Re-ID benchmarks.** Since our generative augmentation can be applied to any Re-ID model, we trained two recent state-of-the-art baselines [5, 22] with Pose-dIVE.

	Human Pose Augmentation	Viewpoint Augmentation	MSMT17		Market-1501		CUHK03 (D)		CUHK03 (L)	
			mAP	R1	mAP	R1	mAP	R1	mAP	R1
(I)	✗	✗	68.0	85.8	89.6	95.5	93.7	95.5	95.5	96.6
(II)	✗	✓	70.9	87.0	90.1	95.3	93.8	95.3	95.8	96.8
(III)	✓	✗	70.9	87.3	90.2	95.4	94.6	96.6	96.4	97.4
(IV)	✓	✓	**71.0**	**87.5**	**90.3**	**95.6**	**95.5**	**97.4**	**97.2**	**97.8**

Table 2. **Quantitative validation of Pose-dIVE augmentation strategies.** We progressively apply human pose and viewpoint augmentation starting from the baseline. The results demonstrate that both types of augmentation independently enhance performance and provide an even greater improvement when combined.

	Training Dataset	# of Images	PIDs	Market1501			
				ResNet-50		SOLIDER	
				mAP ↑	R1 ↑	mAP ↑	R1 ↑
(I)	Baseline Dataset	11,883	619	74.7	88.9	91.6	96.1
(II)	**(I)** + Real Images	30,453	619	77.8	92.8	91.8	96.4
(III)	**(I) + Pose-dIVE Augmented**	30,453	619	**80.2**	**92.9**	**92.3**	**96.6**

Table 3. **Ablation on pose and viewpoint diversity with fixed data size.** Our augmentation strategy achieves better performance gains than simply increasing the dataset with additional in-domain real images.

	Training Dataset	# of Images	PIDs	Market1501			
				ResNet-50		SOLIDER	
				mAP ↑	R1 ↑	mAP ↑	R1 ↑
(I)	Baseline Dataset	11,883	619	74.7	88.9	91.6	96.1
(II)	MARS	495,857	619	72.4	82.3	88.2	89.3
(III)	**(I) + Pose-dIVE Augmented**	30,453	619	**80.2**	**92.9**	**92.3**	**96.6**

Table 4. **Comparison with large-scale in-domain real image augmented dataset.** Increasing the dataset size alone, without ensuring diversity in human poses, does not guarantee improved performance.

formance improvements across various datasets, further validating our augmentation framework and highlighting the efficacy of addressing sparsely distributed human poses and viewpoints.

4.4. Qualitative Results

Qualitative comparisons. In Figure 4, we present a qualitative comparison to a recent GAN-based approach [51]. In contrast to GAN-based methods, our approach effectively generates poses sourced from outside the dataset. GAN-based methods struggle to generalize to diverse poses, often resulting in blurry outputs and a limited ability to maintain the identity of both the reference and the target pose, particularly when dealing with complex features such as accessories, specific clothing details, or bags. This underscores the advantage of utilizing a pre-trained diffusion model, which possesses significant general knowledge about the world.

Visualization of generated data. In Figure 5, we present visualization results on two datasets, MSMT17 [45] and Market-1501 [54]. Given a reference image, our method faithfully preserves its identity while being able to generate diverse poses with high fidelity.

4.5. Ablation Study and Analysis

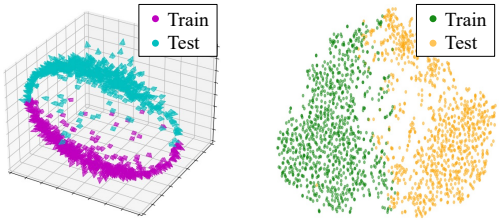
Ablation on the Pose-dIVE augmentation strategy. In Table 2, we conduct an ablation study using CLIP-

Figure 5. **Qualitative results.** Example images from the augmented MSMT17 and Market-1501 dataset demonstrate how the generated images preserve original identities while maintaining realism and consistency with the Re-ID dataset.

ReID [22] to verify the effectiveness of our viewpoint and human pose augmentation. **(I)** serves as the baseline, representing a Re-ID model trained on the original dataset without our augmentation. For **(II)** and **(III)**, we augment viewpoints and human poses, respectively. Both augmentations demonstrate significant performance gains, validating the effectiveness of targeting the biased distributions of human pose and viewpoint for augmentation. **(IV)** demonstrates the full augmentation strategy of our model. The performance gains observed in **(IV)** compared to both **(II)** and **(III)** confirm that both types of augmentation are not only beneficial individually, but also exhibit a complementary effect when combined, leading to further improvements in performance.

Comparison with real image augmented dataset. In this analysis, we investigated whether the performance gain from augmentation originates from an increased dataset size or from increased pose diversity. To validate this, we compared our augmentation method with real-world data collection as shown in Table 3 and Table 4. MARS [55] is an extended version of Market-1501, which samples more images from the same videos used in Market-1501. For simplicity, we conduct experiments using ResNet-50.

In Table 3, the baseline training dataset, **(I)**, is a subset of Market-1501, filtering out identities not present in MARS. **(II)** is the subset of the MARS dataset, matching the number of training images with ours. **(III)** is the dataset augmented with our approach. The performance gap between **(II)** and **(III)**, despite both being trained on datasets of the same size, demonstrates that merely increasing the dataset without introducing diversity in pose and viewpoint results in suboptimal performance.



(a) MSMT17-S (Cam) Viewpoint Distribution (b) MSMT17-S (Pose) Pose Distribution

Figure 6. **Visualization of the split data.** To validate the generalization power of our framework, we split the MSMT17 dataset into train/test sets using two distinct approaches: (a) splitting based on viewpoint, and (b) splitting based on human pose. The visualization clearly illustrates the separation between the train and test distributions.

Method	MSMT17-S (Cam)		MSMT17-S (Pose)	
	mAP ↑	R1 ↑	mAP ↑	R1 ↑
Baseline (SOLIDER [5])	46.6	66.7	60.0	76.0
+ Pose-dIVE	**52.9**	**71.1**	**63.7**	**78.4**

Table 5. **Generalization performance on custom split.** *Cam* denotes dataset split with camera viewpoint, while *Pose* denotes dataset split with human pose. Pose-dIVE demonstrates generalization even on extreme camera viewpoints and human poses not present in the training set.

In Table 4, although the real-image augmented dataset **(II)** is approximately 16 times larger than our augmented dataset **(III)**, the Re-ID model trained on **(II)** performs worse on both metrics compared to the model trained on **(III)**. This highlights that Re-ID model performance is not solely dependent on dataset size but is significantly influenced by the diversity and generalization capability of the images within the dataset.

Analysis on generalization effectiveness. To rigorously assess the generalization capabilities of our framework, we devised an experiment involving an extreme dataset split of the MSMT17 dataset, intentionally amplifying the bias of human pose and viewpoint in the training dataset. We propose two distinct split variants: MSMT17-S (Cam) and MSMT17-S (Pose), where *Cam* divides the dataset into train and test sets with non-overlapping camera viewpoints, while *Pose* divides it based on human pose. These datasets create clear experimental conditions by intentionally partitioning the dataset according to viewpoints and human poses, as visualized in Figure 6. This figure provides a clear visual representation of how the data was split and the lack of overlap between the training and test sets for

Method	Market1501 mAP ↑	Market1501 R1 ↑	Non-pedestrian Dataset mAP ↑	Non-pedestrian Dataset R1 ↑
Baseline (LUPerson [10])	83.0	93.0	75.7	84.2
+ Pose-dIVE	**88.6** (+5.6)	**95.0** (+2.0)	**89.3** (+13.6)	**95.2** (+11.0)

Table 6. **Generalization performance on non-pedestrian test dataset.** We curated a non-pedestrian test dataset consisting of videos that depict unusual human behaviors. The results further validate the robustness of Pose-dIVE in handling diverse human poses.

Method	Market-1501 mAP ↑	Market-1501 R1 ↑
Baseline (CBN [60])	77.3	91.3
+ Pose-dIVE	**82.3**	**93.4**
Baseline (BPBreID [39])	89.4	95.7
+ Pose-dIVE	**89.9**	**95.8**

Table 7. **Pose-dIVE applied to generalization-focused methods.** Re-ID models designed to address camera bias and pose variation benefit from the Pose-dIVE augmented dataset.

both viewpoints and human poses. For each split, we fine-tune the diffusion model on the training split and generate the dataset. Then, we use the generated dataset to train the baseline model. We select SOLIDER as the baseline model for this experiment.

As shown in Table 5, the results clearly demonstrate that our approach significantly improves generalization. The performance enhancements on these filtered datasets were substantially higher, indicating the effectiveness of our method in diversifying the viewpoints and human poses of the original dataset, thereby improving the model's ability to generalize to new, unseen poses, even in this extreme setup. For further details on the dataset split process, please refer to the supplementary material.

Analysis on performance improvements on curated non-pedestrian dataset. We further validate our approach by creating a test dataset that includes a video exhibiting abnormal behaviors, as detailed in Table 6. Unlike traditional pedestrian-centric re-identification datasets, the curated test dataset comprises dynamic images captured in an indoor setting using eight cameras, with an emphasis on abnormal behaviors such as violence and self-harm, which introduce significant diversity in human pose. This dataset includes 23,399 instances across 127 person IDs (PIDs). For our baseline, we employed a ResNet50 model pretrained on LUPerson [10].

As shown in Table 6, our method improved performance on both the existing benchmark dataset Market-1501 and abnormal behaviors CCTV datasets. However, the improvement was significantly larger on our curated test dataset (+13.6 mAP) compared to Market-1501 (+5.6 mAP). This substantial improvement under zero-shot conditions demonstrates the robustness of our method in generalizing to new poses and camera distributions.

Method	Market-1501 mAP ↑	Market-1501 R1 ↑
w/o Pre-trained Diffusion	82.7	93.2
w/ Pre-trained Diffusion	**90.3**	**95.6**

(a)

Diffusion model trained on	Market1501 mAP ↑	Market1501 R1 ↑
In-domain pose and viewpoint	89.6	95.2
External pose and viewpoint	**90.3**	**95.6**

(b)

Table 8. **Ablation Studies.** (a) The effect of using a pre-trained diffusion model for weight initialization. (b) A comparison between diffusion models trained on in-domain versus external pose and viewpoint data. For both studies, we train the CLIP-reID model [22] on the generated dataset.

Analysis on the synergy of generalization-focused Re-ID models with Pose-dIVE. In Table 7, we explore the synergy between our approach and Re-ID models specifically designed to enhance generalization, particularly those addressing variations in camera and human poses.

We evaluate the impact of our approach on two representative models: CBN [60], which addresses camera bias, and BPBreID [39], designed to handle human pose variations. We select these models due to their public availability and proven effectiveness in their domains. While both models are already effective, they show marked improvements when trained on the dataset augmented by our approach. These results highlight that even with algorithmic advancements in pose generalization, there remains significant potential for improvement through dataset enhancements.

Ablation on pre-trained diffusion model. In Table 8 (a), we conduct an ablation study to evaluate the impact of utilizing the pre-trained Stable Diffusion model [37] for generating diverse camera viewpoints and human poses in our augmented dataset. Employing the pre-trained diffusion model results in a notable 7.6% mAP improvement, underscoring the significance of foundation models in the context of dataset augmentation.

Ablation on pose and viewpoint diversification. Table 8 (b) presents a comparison between a model trained on generated data using in-domain poses and viewpoints and one trained using external poses and viewpoints. The model trained with pose-diversified data demonstrates a notable performance improvement, which underscores the effectiveness of our proposed method.

5. Conclusion

In this paper, we proposed Pose-dIVE, a novel data augmentation approach that leverages pre-trained diffusion models to diversify pose and viewpoint distributions in person re-identification training datasets. The key to success of Pose-dIVE lies in diversifying human pose and viewpoint by integrating pose, viewpoint, and identity conditions into large-scale pre-trained diffusion models, effectively leveraging the vast knowledge embedded in these models to generate high-quality augmented data. Comprehensive experiments demonstrated the effectiveness of our approach, with Pose-dIVE achieving significant performance improvements compared to baselines.

Acknowledgment This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279, RS-2025-II212068, RS-2023-00227592, RS-2025-02214479, RS-2024-00457882, RS-2025-25441838, RS-2025-25441838, RS-2025-02214479, RS-2025-02217259) and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2024-00333068, RS-2023-00222280, RS-2023-00266509), and National Research Foundation of Korea (RS-2024-00346597).

References

[1] Slawomir Bak, Sofia Zaidenberg, Bernard Boulay, and François Bremond. Improving person re-identification by viewpoint cues. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 175–180. IEEE, 2014. 1

[2] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019. 4, 5

[3] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2004–2013, 2021. 1, 2, 3

[4] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Learning invariance from generated variance for unsupervised person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7494–7508, 2022. 1, 2, 3

[5] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15050–15061, 2023. 1, 2, 5, 6, 7

[6] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3300–3310, 2020. 6

[7] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 40 (2), 2018. 1

[8] Yeong-Jun Cho and Kuk-Jin Yoon. Improving person re-identification via pose-aware multi-shot matching. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 1354–1362. IEEE Computer Society and the Computer Vision Foundation (CVF), 2016. 1

[9] Changxing Ding, Kan Wang, Pengfei Wang, and Dacheng Tao. Multi-task learning with coarse priors for robust part-aware person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1474–1488, 2020. 6

[10] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14750–14759, 2021. 1, 8

[11] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in Neural Information Processing Systems*, pages 1229–1240, 2018. 1, 3

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[13] Jianyang Gu, Kai Wang, Hao Luo, Chen Chen, Wei Jiang, Yuqiang Fang, Shanghang Zhang, Yang You, and Jian Zhao. Msinet: Twins contrastive search of multi-scale interaction for object reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19243–19253, 2023. 6

[14] 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. 6

[15] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 2, 4, 5

[16] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5098–5107. IEEE Computer Society, 2018. 2

[17] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11173–11180, 2020. 6

[18] Srikrishna Karanam, Yang Li, and Richard J Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4516–4524. IEEE, 2015. 1

[19] Arnab Karmakar and Deepak Mishra. Pose invariant person re-identification using robust pose-transformation gan. *arXiv preprint arXiv:2105.00930*, 2021. 1

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from

equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2288–2295. IEEE, 2012. 2

[22] 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. 1, 2, 5, 6, 7, 8

[23] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 1, 5

[24] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3685–3693, 2015. 2

[25] Feng Liu, Minchul Kim, ZiAng Gu, Anil Jain, and Xiaoming Liu. Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19617–19626, 2023. 1, 3

[26] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018. 1, 2

[27] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3550–3557. IEEE Computer Society, 2014. 1

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *Acm Transactions on Graphics*, 34 (Article 248), 2015. 2, 3

[29] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019. 2

[30] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Data-augmentation for reducing dataset bias in person re-identification. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE Computer Society, 2015. 1

[31] Xingyang Ni and Esa Rahtu. Flipreid: closing the gap between training and inference in person re-identification. In *2021 9th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2021. 6

[32] 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. 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)*, 2018. 1

[34] Alec Radford. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[36] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1025–1034, 2021. 6

[37] 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, 4, 5, 8

[38] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 420–429, 2018. 1

[39] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1613–1623, 2023. 2, 8

[40] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 717–734. Springer, 2020. 3

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 2

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[43] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2540–2549, 2022. 6

[44] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013. 1

[45] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 1, 5, 6

[46] Zexian Yang, Dayan Wu, Chenming Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. A pedestrian is worth one prompt: Towards language guidance person re-identification. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17343–17353, 2024. 1

[47] 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. 6

[48] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 5

[49] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 5

[50] Xianghao Zang, Ge Li, Wei Gao, and Xiujun Shu. Learning to disentangle scenes for person re-identification. *Image and Vision Computing*, 116:104330, 2021. 6

[51] 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*, pages 7713–7722, 2022. 5, 6

[52] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 1

[53] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1

[54] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 1, 2, 5, 6

[55] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 868–884. Springer, 2016. 1, 7

[56] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[57] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1224–1244, 2017. 1

[58] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. 1, 2

[59] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2138–2147, 2019. 2

[60] Z Zhong, L Zheng, Z Zheng, S Li, and Y Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. 8

[61] 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. 1

[62] Haidong Zhu, Pranav Budhwant, Zhaoheng Zheng, and Ram Nevatia. Seas: Shape-aligned supervision for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 164–174, 2024. 1

[63] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 6