Diverse Person: Customize Your Own Dataset for Text-Based Person Search

Zifan Song[1], **Guosheng Hu**[2], **Cairong Zhao**[1]*

[1]Department of Computer Science and Technology, Tongji University, China
[2]Oosto, Belfast, U.K., BT1 2BE
{sugger, zhaocairong}@tongji.edu.cn, huguosheng100@gmail.com

Abstract

Text-based person search is a challenging task aimed at locating specific target pedestrians through text descriptions. Recent advancements have been made in this field, but there remains a deficiency in datasets tailored for text-based person search. The creation of new, real-world datasets is hindered by concerns such as the risk of pedestrian privacy leakage and the substantial costs of annotation. In this paper, we introduce a framework, named Diverse Person (DP), to achieve efficient and high-quality text-based person search data generation without involving privacy concerns. Specifically, we propose to leverage available images of clothing and accessories as reference attribute images to edit the original dataset images through diffusion models. Additionally, we employ a Large Language Model (LLM) to produce annotations that are both high in quality and stylistically consistent with those found in real-world datasets. Extensive experimental results demonstrate that the baseline models trained with our DP can achieve new state-of-the-art results on three public datasets, with performance improvements up to 4.82%, 2.15%, and 2.28% on CUHK-PEDES, ICFG-PEDES, and RSTPReid in terms of Rank-1 accuracy, respectively.

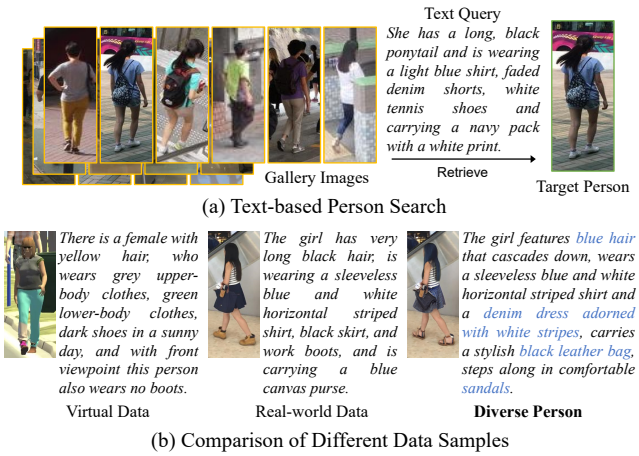


Figure 1: (a) Illustration of the text-based person search task. The yellow and green bounding boxes denote the gallery persons and the pedestrian correctly matched, respectively. (b) Comparisons among the virtual data (left), real-world data (middle), and the proposed Diverse Person (right). We highlight the modified attributes in the text description.

Introduction

Person re-identification (Re-ID) aims to retrieve specific pedestrians from a gallery set collected across cameras. Over the past decade, Re-ID has achieved notable advancements and expanded into various specialized subtasks. Based on the type of query data, Re-ID can be roughly divided into four subtasks: traditional Re-ID (Yang et al. 2017; Liu et al. 2022) based on cropped images, video Re-ID (Yan et al. 2018; Hou et al. 2021) based on video data, person search (Xu et al. 2014; Song et al. 2023) based on scene images, and text-based person search (Li et al. 2017; Shu et al. 2022; Jiang and Ye 2023) based on textual descriptions. Among these, text-based person search, distinguished by its use of straightforward and widely accessible text descriptions to locate target subjects, has seen a surge in interest due to its practical applications.

Involving information processing from two heterogeneous modalities, text-based person search encounters several challenges, including the potential for query text misinterpretation and diminished accuracy due to limited information. To track these challenges and enhance the quality of search results, cross-modal alignment along with extensive text-image datasets, is necessary. By creating loss functions or network models, early techniques (Li et al. 2017; Zhang and Lu 2018; Chen et al. 2021) align images and texts within a unified embedding space. However, these approaches ignore local fine-grained alignment, focusing instead on global image-text correlations. As a result, recent researches (Wang et al. 2020; Zhu et al. 2021; Wu et al. 2021) shift towards modeling the nuanced, local correspondences between modalities. This typically involves segmenting images and texts into local components and then establishing correspondence between them. The effectiveness of these methods is contingent upon the quality of the local components. In order to further supplement clues available for matching, some previous methods adopt external models to extract useful information (Niu et al. 2020; Zhu et al. 2021; Wu et al. 2021). Specially, Niu et al. (2020) introduce

*Corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

a hierarchical multi-granularity image-text alignment mechanism and the cross-granularity mapping is constructed using additional alignment between the global image and noun phrases, as well as between the entire sentence and the horizontal image stripes.

Currently, improving cross-modal alignment techniques to enhance retrieval performance has reached a bottleneck, while concurrently, there remains a scarcity of text-based person search datasets. Collecting new real-world datasets presents problems such as pedestrian privacy leakage and high annotation costs. Some existing methods (Wang, Liang, and Liao 2022; Zhang et al. 2021; Wang, Liao, and Shao 2020) attempt to address these issues by exploring the use of virtual datasets. However, as shown in Fig. 1 (b), virtual datasets still suffer from significant domain gaps, and the generated text annotations have limited diversity compared to descriptions written by human annotators.

In this paper, we strive to resolve the aforementioned challenges by introducing a novel framework, Diverse Person (DP). Specifically, we tactfully propose a method that leverages available images of clothing and accessories in the form of reference attribute images. These images are utilized to augment the original dataset images via diffusion models. Subsequently, we employ a Large Language Model (LLM), based on the original annotations and incorporating the attributes from the reference attribute images, to generate high-quality annotations stylistically consistent with those found in the existing real-world dataset. As shown in Fig. 1 (c), we achieve efficient and high-quality text-based person search data generation while circumventing additional privacy concerns. Furthermore, by incorporating semantic features in the structure of the diffusion model, we guide the model's attention towards pedestrian attributes, effectively improving the model's discriminative ability for different pedestrian attributes and achieving stronger performance.

Our main contributions are summarized as follows:

- We introduce an innovative framework, Diverse Person (DP), specifically designed for the efficient generation of high-quality text-based person search datasets without involving privacy leakage issues.
- We propose to edit both images and textual content of the original data based on pedestrian reference attributes via diffusion models and a Large Language Model (LLM), achieving more diverse and authentic generative effects.
- Extensive experimental results demonstrate the universality and efficacy of our method to further the performance of the existing state-of-the-art ones on three popular text-based person search datasets CUHK-PEDES, RSTPReid, and ICFG-PEDES.

Related Work

Text-based Person Search

Text-based person search aims to locate the target pedestrian in gallery images based on a given query text. The earliest definition of the text-based person search task dates back to 2017: Li et al. (2017) propose the first benchmark dataset for text-based person search, CUHK-PEDES. The primary challenge of text-based person search is how to achieve high-quality cross-modal alignment for fast search. To track this challenge, early researches (Li et al. 2017; Zhang and Lu 2018) primarily focus on dealing with global text-image matching and utilize matching losses for cross-modal alignment. These global matching-based algorithms are straightforward and effective, however, they may neglect essential local features and inadvertently incorporate noisy data. More recent research (Niu et al. 2020; Niu, Huang, and Wang 2020; Ding et al. 2021; Wang et al. 2022a) has widely employed additional local feature learning branches, which leverage information such as human body segmentation and clothing color to explore local matches between images and text. While it has been demonstrated that these methods offer superior retrieval results when compared to employing only global features, they also add more computational complexity to inference when computing image-text similarity.

All of the aforementioned research extract visual and textual characteristics using backbones that have been individually trained with uni-modal data, and they subsequently execute cross-modal alignment without taking advantage of the excellent cross-modal alignment capabilities of recently promising vision-language pre-training models. To this end, some state-of-the-art methods have utilized Transformers (Shao et al. 2022) and cross-modal pre-training models like CLIP (Radford et al. 2021a) to achieve superior performance. Han et al. (2021) first introduce a CLIP model for text-to-image person retrieval, using a momentum contrastive learning framework to transfer knowledge learned from large-scale general image-text pairs. Subsequently, Yan et al. (2023a) propose a CLIP-driven fine-grained information mining framework to propagate the knowledge from CLIP. Jiang et al. (2023) introduce IRRA to learn more discriminative image-text embeddings.

Enlarging Training Data with Generative Models

A lot of research has been done on the use of generative models for supplementing training data(Huang et al. 2018; Zhou et al. 2022). These techniques either pre-train GANs or use them to produce images from the desired distribution. In the person search task, Yao et al. (2020) propose to implement GAN to generate unlabeled samples, which helps to improve the discriminating ability of the model. According to recent research (He et al. 2022; Shipard et al. 2023; Zhu et al. 2023), diffusion models (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Song and Ermon 2019) can produce training data in zero-shot or few-shot conditions as well as challenging training examples, and have achieved state-of-the-art results on text-to-image generation (Rombach et al. 2022; Saharia et al. 2022). These studies indeed demonstrate the potential of diffusion-generated data, but they stay more in the qualitative analysis (*e.g.*, visualization of the generated image data). In this paper, our approach adopts diffusion models for text-to-image editing instead of solely generating images from text, resulting in obtaining more realistic pedestrian images that are consistent with the style of the source data.

Methodology

In this section, we first describe the problem formulation of texted-based person search and present preliminary on diffusion model. Then, we introduce our proposed framework Diverse Person (DP) in detail.

Problem Formulation

As illustrated in Fig. 1 (a), the objective of text-based person search is to identify the most closely matching pedestrian in a set of gallery images $\mathcal{G} = \{g_1, g_2, ..., g_N\}$ corresponding to a given query text q. To achieve superior matching performance, we propose to supplement the training set by customizing text-based person search data using diffusion models and improve the discriminative ability for different pedestrian attributes of text-based person search models.

Preliminary on Diffusion Models

Diffusion models, a category of probabilistic generative models, are designed to discover data distributions by progressively reducing noise from randomly sampled Gaussian distributions. Theoretically, this process involves learning the inverse of a T-length fixed Markov Chain.

Given a set of image-text pairs, *i.e.*, $\mathcal{D}_{\text{train}} = \{(\mathcal{I}_1, \mathcal{T}_1), \ldots, (\mathcal{I}_N, \mathcal{T}_N)\}$, the text-to-image synthesis model functions as a series of conditional denoising neural networks with equal weighting. Based on the text prompt $\epsilon_\theta(\mathcal{I}_i^t, t, \mathcal{T}_i)$, it iteratively predicts a denoised variant of the input at each timestep, where $\mathcal{I}_i^t$ represents a noisy version of the input image, $t \in \{1, \ldots, T\}$ indicates the timestep, and $i \in \{1, \ldots, N\}$. The Stable Diffusion (Rombach et al. 2022) variation of the diffusion model is one that we specifically draw upon. This variant transitions the diffusion mechanism to latent space, employing a variational autoencoder to encode images. Below, we provide a concise overview of its architecture and training methodology..

Architecture Stable Diffusion is comprised of three key components: a text encoder for generating textual embeddings; a pre-trained variational autoencoder (VAE) tasked with encoding and decoding image-related latent vectors; and a time-conditional UNet, denoted as $\epsilon_\theta(\cdot)$, which progressively denoises these latent vectors. The denoising process is facilitated by a series of convolutional operations, involving the downsampling and upsampling of visual feature maps connected by skip connections. Crucially, the integration of visual and textual information occurs within the UNet, where the embeddings of both modalities interact via cross-attention layers. To elaborate, the process commences with the text encoder projects the textual prompt $\mathcal{T}$ into textual embeddings. These embeddings are subsequently transformed into Key and Value components. Concurrently, the spatial feature map of the noisy image is linearly projected into the Query component. Through iterative attention mechanisms conditioned on the textual prompt, the Query component is updated, leading to the gradual denoising of the latent vectors.

Training and Inference The following delineates the training methodology for Stable Diffusion: Initially, an input image is transformed into a latent vector z, based on a training pair $(\mathcal{I}, \mathcal{T})$. During this process, a variably-noised vector $z^t := \alpha^t z + \sigma^t \epsilon$ is generated, where α^t and σ^t are parameters governing the noise schedule and sample quality, and $\epsilon \sim \mathcal{N}(0, 1)$ is a noise term. The training strategy focuses on predicting the noise ϵ and reconstructing the original vector z by conditioning on the text prompt $\mathcal{T}$. This is achieved through the optimization of a time-conditional UNet. The training employs a squared error loss function for the expected noise term, formalized as:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t,\mathcal{T}}\left[\left\|\epsilon - \epsilon_\theta\left(z^t, t, \mathcal{T}\right)\right\|_2^2\right] \quad (1)$$

where t is sampled from $\{1, \ldots, T\}$. During inference, Stable Diffusion is sampled by iteratively denoising $z^T \sim \mathcal{N}(0, 1)$ conditioned on the text prompt $\mathcal{T}$. Specifically, at each denoising step $t \in \{1, \ldots, T\}$, z^{t-1} is obtained from both z^t and the predicted noise term of the UNet, which inputs z^t and the text prompt $\mathcal{T}$. After the final denoising step, z^0 is mapped back to generate the resulting image $\mathcal{I}$.

Diverse Person

Attribute Embedding for Text Representation We suggest the integration of visual cues derived from specific attribute images into language prompts, aiming to facilitate image generation that accurately reflects pedestrian attributes without requiring tuning. This method involves three key elements: a set of reference attribute images, denoted as $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$; a text prompt, represented as $\mathcal{T} = \{w_1, w_2, \ldots, w_n\}$; and a look-up list $L = \{l_1, l_2, \ldots, l_m\}$, where each $l_i \in \{1, 2, \ldots, n\}$ indicates the correspondence between a specific attribute and a word in the text prompt. The process begins with the embedding of the text prompt $\mathcal{T}$ and the reference attribute $\mathcal{A}$ using the CLIP text and image encoders, respectively. Subsequently, the word embeddings are enriched with visual information extracted from the reference attributes through a multilayer perceptron (MLP). This enhancement results in a fusion of the word embeddings with the visual features. These augmented embeddings are then input into the MLP for further processing. The final conditioning embeddings produced by above, are designated as $\hat{c}$, are as follows:

$$\hat{c}_k = \begin{cases} \psi(\mathcal{T})_k, & k \notin L \\ \text{MLP}\left(\psi(\mathcal{T})_k \| \phi\left(a_i\right)\right), & k = l_i \in L \end{cases} \quad (2)$$

where $\hat{c}_k$ represents the augmented conditional embeddings of the k-th word, $\psi(\mathcal{T})_k$ and $\phi\left(a_i\right)$ stand for the embedding of the k-th word from text prompt $\mathcal{T}$ and attribute image a_i encoded by the pre-trained CLIP text and image encoders ψ and ϕ, respectively.

Attribute-Driven Pedestrian Image Customization To facilitate inference-only attribute-driven pedestrian image customization (generation) that relies solely on inference, we align noun phrases from common image text descriptions with corresponding attribute segments present in target images. We first aggregate all noun words (such as "hair") in the image captions using a dependency parsing model, and all attributes in the images are then segmented

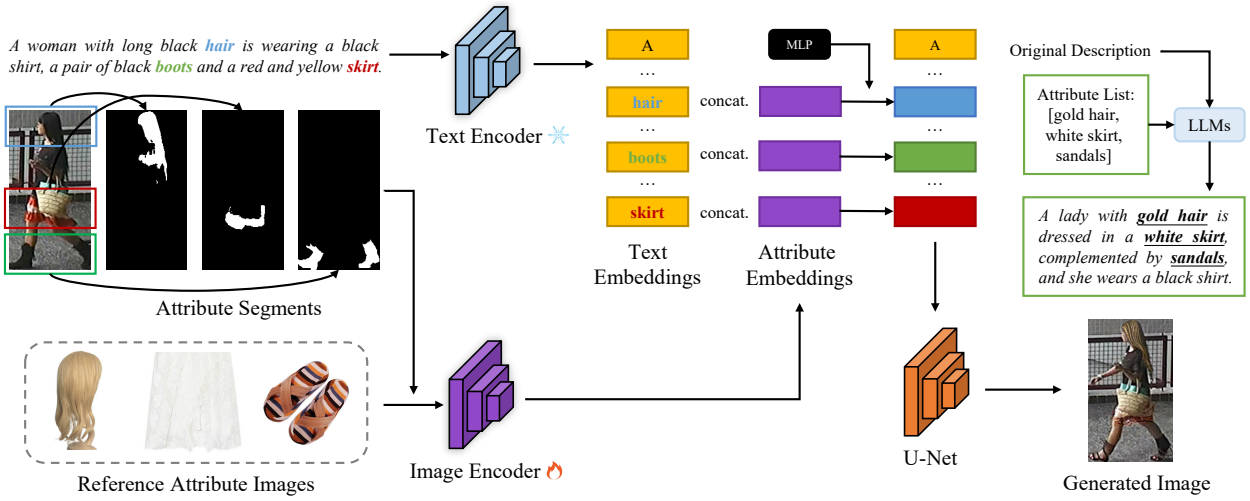


Figure 2: The overall framework of the proposed Diverse Person (DP) for high-quality text-based person search data generation.

using a panoptic segmentation model (*e.g.*, SAM (Kirillov et al. 2023)). Then, we couple these attribute segments with the appropriate noun phrases in the captions using a greedy matching method based on text and image similarity (Reimers and Gurevych 2019; Radford et al. 2021b). Eq. (2) details the use of attribute-enhanced conditioning to denoise the disturbed target image during training. To prevent overfitting to attribute backgrounds, we apply random noise to mask the attribute backgrounds prior to encoding. As a result, our DP does not need to explicitly separate the backdrop when using natural attribute images during inference. According to Fig. 2, we train the image encoder, MLP module, and U-Net using the denoising loss.

Using enhanced text representation in place of initial text features for inference often results in visuals that are highly similar to the reference attributes, enabling the generation of customized pedestrian samples that are highly consistent with the reference attribute images. This is particularly useful for strict customization of colors and textures such as hair and clothing when there are sufficient reference attribute images. To provide better diversity when there are fewer available reference attribute images, we apply a delayed conditioning method (Xiao et al. 2023) below, allowing the model to balance diversity with consistency in reference attributes:

$$\epsilon^t = \begin{cases} \epsilon_\theta\left(z^t, t, c\right) & \text{if } t > \gamma T \\ \epsilon_\theta\left(z^t, t, \hat{c}\right) & \text{otherwise} \end{cases} \tag{3}$$

where the text embeddings c and $\hat{c}$ stand for the original text embedding and the text embedding enhanced with the input image embedding, respectively. γ is a hyperparameter that shows how much attribute conditioning there is.

Regularization with Attribute Segmentation Masks When customizing multiple attributes in a pedestrian image, models without regularization often exhibit overlapping influences from multiple reference attributes within the same region of generation. This overlap can lead to incongruities in the pedestrian's appearance, diverging from the intended reference attribute. For instance, a pedestrian intended to be customized as wearing jeans and carrying a leather bag may be inaccurately rendered with a composite "denim-leather" bag. According to earlier researches (Hertz et al. 2022; Xiao et al. 2023), the cross-attention mechanism observed in diffusion models governs the layout of generated images. To prevent the fusion of attribute appearances, the attention map for each attribute token should focus primarily on the area of the corresponding attribute, rather than the entire image, *i.e.*, appropriate cross-attention maps should approximate segmentation of the target attribute, clearly distinguishing features related to different attributes.

To achieve this, we add a regularization term to the attribute cross-attention maps during training, promoting focused attention on specific attribute areas. Specifically, we employ the following regularization loss to minimize the distance between the segmentation mask and the cross-attention map of the corresponding attribute:

$$\mathcal{L}_{\text{regularization}} = \frac{1}{m}\sum_{i=1}^{m} |A_{l_i} - s_i|_{\text{avg}} \tag{4}$$

$$\mathcal{L} = \mathcal{L}_{\text{diffusion}} + \lambda \mathcal{L}_{\text{regularization}} \tag{5}$$

where s_i indicates the segmentation mask of the i-th attribute to be customized, A_{l_i} represents the cross-attention map of the l_i-th word corresponding to the i-th attribute, and $|A_{l_i} - s_i|_{\text{avg}}$ denotes the average L1 distance (mean absolute difference) between the cross-attention map A_{l_i} and the segmentation mask s_i.

Annotating with Large Language Models (LLMs) To obtain accurate and linguistically diverse textual annotations, we experiment with ChatGPT, a large language model (LLM) that has demonstrated remarkable results in natural language processing tasks. For synthetic data that can augment real data, we presume that the generated descriptions, which are more similar to the real data, have higher quality. To this end, we have optimized more than a dozen prompts

I: {The girl has very long black hair, is wearing a sleeveless blue and white horizontal striped shirt, black skirt, and work boots, and is carrying a blue canvas purse.} {blue hair, Denim dress with white stripes, black leather bag, sandals} **O:** The girl features blue hair that cascades down, wears a sleeveless blue and white horizontal striped shirt and a denim dress adorned with white stripes, carries a stylish black leather bag, steps along in comfortable sandals.
I: {A dark haired person wearing a red short sleeved shirt and black pants with white shoes.} {white hair, green shorts, black umbrella} **O:** A person sporting white hair is attired in green shorts, along with a black umbrella and white shoes.

Table 1: Two exemplars in our prompt for annotation generating using a Large Language Model (LLM).

based on real descriptions to generate textual annotations and the final version of the prompt template is as follows:

"You are asked to rewrite a sentence describing pedestrians substituting the corresponding attribute (refer to the attribute list). The sentence to be rewritten is as follows: {original sentence} The attribute list: {attribute list} The rewritten sentence:"

In order to illustrate the effectiveness of our prompt more clearly, we provide two specific exemplars in Tab. 1. It should be noted that, for the sake of conciseness and clarity, the input only displays the dynamic content excluding the template.

Experiments

In this section, we conduct extensive experiments on three datasets to verify the superiority of our DP framework. Furthermore, we present analytical studies to provide a more in-depth demonstration of the efficacy of our method.

Experiment Settings

Datasets We access three widely used benchmark datasets for text-based person search to evaluate our approaches.

CUHK-PEDES (Li et al. 2017) comprises 40,206 images and 80,412 text descriptions of 13,003 persons. Each image is manually annotated with 2 text descriptions with an average length of not less than 23 words. The database constitutes a vocabulary with 9408 unique words. To enable a fair comparison with existing methods, we follow the official data split protocol in (Li et al. 2017). Specifically, 34,054 images of 11,003 persons and corresponding 68,108 text descriptions are used as the training set. The remaining 2000 persons are equally divided into the validation and testing sets, with the validation set comprising 3,078 images and 6,156 text descriptions, and the testing set comprising 3,074 images and 6,148 text descriptions.

ICFG-PEDES (Ding et al. 2021) contains 54,522 images of 4,102 persons collected from the MSMT17 (Wei et al. 2018) database, with each image having a corresponding text description of an average length of 37 words. The vocabulary contains 5554 unique words. Following the data split protocol in (Ding et al. 2021), the database is divided into training and testing sets. The training set contains 34674 image-text pairs of 3102 persons, while the testing set consists of 19848 image-text pairs of the remaining 1000 persons.

RSTPReid (Zhu et al. 2021) consists of 4,101 identities, with each identity having 5 corresponding images, resulting in a total of 20,505 images. These images are captured by 15 different cameras, and each image is accompanied by 2 textual descriptions. For the dataset split, the training set, validation set, and test set contain 3,701, 200, and 200 identities, respectively.

Evaluation Protocols We utilize the same evaluation protocols as in previous works (Jiang and Ye 2023), where the mean Average Precision (mAP) and the Rank-k metrics (k=1,5,10) are adopted as evaluation metrics. The Rank-k metrics represents the probability of finding at least one matching pedestrian within the top-k candidate list when a textual description is used as a query. Higher Rank-k and mAP values indicate better performance.

Implementation Details We use PyTorch (Paszke et al. 2017) to implement and train all the corresponding models in this paper. During the training of the diffusion models, we use the pre-trained Stable Diffusion v1-5 and conduct model training for 30k steps on 2 NVIDIA RTX 3090 GPUs. We set a maximum of 3 reference attributes with a learning rate of 1e-5 and a batch size of 32. The regularization loss is employed to the downsampled cross-attention maps and the hyperparameter λ is set to 0.001. For the hyperparameter γ, we set it to 0.73 ([0.7, 0.75] also achieve optimal results), to achieve a balance between diversity and consistency with the reference attributes. In the training phase of text-based person search models, our DP functions as a form of data augmentation, while in the testing phase, we adhere to the inference steps of the corresponding baseline models without requiring extra knowledge. For the image encoder ResNet-50 (He et al. 2016), we adopt minor modifications: remove the average pooling layer and the fully connected layer, and set the stride of the last convolution layer to 1 for larger feature map. All images are resized to 384×128 before being fed into the image encoders, followed by random horizontally flipping, random crop with padding, and random erasing for data augmentations. For the text encoder BERT (Vaswani et al. 2017), the BERT encoder is frozen and the dimensions of textual features are 768. The sizes of the vocabulary vary for different databases, with CUHK-PEDES set to 5000 and ICFG-PEDES set to 3000.

Comparisons with State-of-the-art Methods

In this section, we compare with the state-of-the-art models on CUHK-PEDES, ICFG-PEDES, and RSTPReid.

Performance Comparisons on CUHK-PEDES The experimental results of Tab. 2 demonstrate the superiority of our method in terms of Rank-k and mAP accuracy over other competitors on CUHK-PEDES. Our proposed method achieves impressive performance of 75.66% Rank-1 accuracy. Compared to our baseline SRCF (Suo et al. 2022),

Method	Publication	Image Encoder	Text Encoder	Additional Knowledge	R1	R5	R10	mAP
GNA-RNN (Li et al. 2017)	CVPR'17	ResNet-50	LSTM	None	19.05	-	53.64	-
GLA (Chen et al. 2018)	ECCV'18	ResNet-50	LSTM	None	43.58	66.93	76.20	-
CMPM (Zhang and Lu 2018)	ECCV'18	ResNet-50	LSTM	None	49.37	-	79.27	-
ViTAA (Wang et al. 2020)	ECCV'20	ResNet-50	LSTM	Segmentation	54.92	75.18	82.90	51.60
TDE (Niu, Huang, and Wang 2020)	ACMMM'20	ResNet-50	BERT	Dependency	56.67	76.85	84.63	-
DSSL (Zhu et al. 2021)	ACMMM'21	ResNet-50	BERT	Surroundings	59.98	80.41	87.56	-
SSAN (Ding et al. 2021)	arXiv'21	ResNet-50	LSTM	None	61.37	80.15	86.73	-
LapsCore (Wu et al. 2021)	ICCV'21	ResNet-50	BERT	Colorization	63.40	-	87.80	-
ISANet (Yan et al. 2023b)	TNNLS'23	ResNet-50	LSTM	None	63.92	82.15	87.69	-
LBUL (Wang et al. 2022b)	ACMMM'22	ResNet-50	BERT	None	64.04	82.66	87.22	-
CAIBC (Wang et al. 2022a)	ACMMM'22	ResNet-50	BERT	Colorization	64.43	82.87	88.37	-
AXM-Net (Farooq et al. 2022)	AAAI'22	ResNet-50	BERT	None	64.44	80.52	86.77	58.73
SRCF (Suo et al. 2022)	ECCV'22	ResNet-50	BERT	None	64.88	83.02	88.56	-
LGUR (Shao et al. 2022)	ACMMM'22	DeiT-S	BERT	None	65.25	83.12	89.00	-
IVT (Shu et al. 2022)	ECCV'22	ViT-B	BERT	Augmentation	65.59	83.11	89.21	-
CFine (Yan et al. 2023a)	TIP'23	CLIP-ViT	BERT	None	69.57	85.93	91.15	-
IRRA (Jiang and Ye 2023)	CVPR'23	CLIP-ViT	CLIP-Xformer	None	73.38	89.93	93.71	66.13
SRCF+DP (ours)	Proposed	ResNet-50	BERT	Augmentation	69.13	86.24	90.32	-
IVT+DP (ours)	Proposed	ViT-B	BERT	Augmentation	70.41	88.75	91.18	-
IRRA+DP (ours)	Proposed	CLIP-ViT	CLIP-Xformer	Augmentation	**75.66**	**90.59**	**94.07**	**66.58**

Table 2: Performance comparisons with state-of-the-art methods on the CUHK-PEDES dataset.

Method	R1	R5	R10	mAP
Dual Path (Zheng et al. 2020)	38.99	59.44	68.41	-
CMPM (Zhang and Lu 2018)	43.51	65.44	74.26	-
ViTAA (Wang et al. 2020)	50.98	68.79	75.78	-
SSAN (Ding et al. 2021)	54.23	72.63	79.53	-
IVT (Shu et al. 2022)	56.04	73.60	80.22	-
LGUR (Shao et al. 2022)	59.02	75.32	81.56	-
CFine (Yan et al. 2023a)	60.83	76.55	82.42	-
IRRA (Jiang and Ye 2023)	63.46	80.25	85.82	38.06
IRRA+DP (ours)	**65.61**	**81.73**	**86.95**	**39.14**

Table 3: Performance comparisons with state-of-the-art methods on the ICFG-PEDES dataset.

Method	R1	R5	R10	mAP
DSSL (Zhu et al. 2021)	39.05	62.60	73.95	-
SSAN (Ding et al. 2021)	43.50	67.80	77.15	-
LBUL (Wang et al. 2022b)	45.55	68.20	77.85	-
IVT (Shu et al. 2022)	46.70	70.00	78.80	-
CAIBC (Wang et al. 2022a)	47.35	69.55	79.00	-
CFine (Yan et al. 2023a)	50.55	72.50	81.60	-
IRRA (Jiang and Ye 2023)	60.20	81.30	88.20	47.17
IRRA+DP (ours)	**62.48**	**83.77**	**89.93**	**48.86**

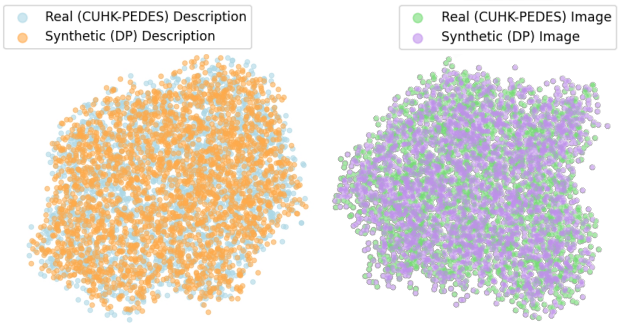
Table 4: Performance comparisons with state-of-the-art methods on the RSTPReid dataset.

our method yields a 4.25%, 3.22%, and 1.76% improvement in Rank-1, Rank-5, and Rank-10, respectively. When combined with the recent state-of-the-art model IRRA (Jiang and Ye 2023), our proposed DP can further exceed the performance by 2.28%, 0.66%, 0.36%, and 0.45% in terms of Rank-1, Rank-5, Rank-10, and mAP, respectively.

Performance Comparisons on ICFG-PEDES We also conduct performance experiments on ICFG-PEDES and Tab. 3 reports that our DP further boosts the baseline IRRA by 2.15%, 1.48%, 1.13%, and 1.08% in terms of Rank-1, Rank-5, Rank-10, and mAP. Additionally, our approach surpasses the recent state-of-the-art model CFine (Yan et al. 2023a) by 4.78%, 5.18%, 4.53% on all Rank-k metrics.

Performance Comparisons on RSTPReid In addition, we have accessed the newly released large-scale database RSTPReid to prove the generality of the proposed DP. Since this database is up-to-date, only a few methods are available for comparison. As shown in Tab. 4, when combined with IRRA, our IRRA+DP outperforms all existing methods reported on RSTPReid in all metrics, achieving promising performance of 62.48%, 83.77%, 89.93%, and 48.86% on Rank-1, Rank-5, Rank-10, and mAP, respectively. Compared to the recent approach CFine (Yan et al. 2023a), our method achieves a remarkable performance margin of 11.93%, 11.27%, 8.33% w.r.t Rank-1, Rank-5, and Rank-10, respectively.

Universality of DP and CTGM As a general approach, our DP can be applied in a plug-and-play manner to enhance the performance of other methods. We have apply our approach to three frameworks including SRCF (Suo et al. 2022), IVT (Shu et al. 2022), and IRRA (Jiang and Ye 2023), evaluated on three popular benchmark datasets. As shown in Tab. 2, our method can effectively enhance the performance of SRCF, IVT, and IRRA, in particular, the improvements of IVT is significant ($4.82\% \uparrow$, $5.64\% \uparrow$, and $1.97\% \uparrow$ in terms of Rank-1, Rank-5, and Rank-10, respectively). In



(a) Textual features extracted by CLIP-Xformer. (b) Visual features extracted by CLIP-ViT.

Figure 3: Distribution analysis via comparing T-SNE plot clusters of real and customized (synthetic) samples.

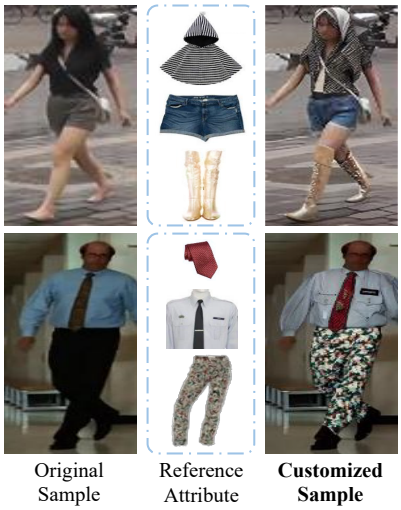


Figure 4: Customized examples generated by DP and the reference attributes are highlighted in the text descriptions.

addition, these baselines involve three different combinations of image encoders and text encoders (*i.e.*, ResNet-50 with BERT, ViT-B with BERT, and CLIP-ViT with CLIP-Xformer), further verifying the effectiveness and universality of the method we proposed.

Analytical Study

Quality Analysis of the Synthetic Data For both images and text, our DP does not generate synthetic data from scratch but rather edits certain attributes within real data, resulting in a smaller synthetic-to-real gap. To further valid the quality of the customized samples, we visualize the feature space of the real data with our synthetic data using t-SNE (Linderman et al. 2017) plots in Fig. 3. Clearly, the features of the real and synthetic data are intermingled and lack of separation, indicating that the high-quality samples from DP are capable of diversifying and augmenting the existing datasets.

Qualitalive Results We provide examples of customized (synthetic) results generate by DP in Fig. 4 and visualize the retrieving results in Fig. 5. The Rank-5 text-based person search matches on CUHK-PEDES are reported. Compared with the baseline method, our approach is more robust in distinguishing similar-looking pedestrians, leading to the best performance with the correct matching ranked at the top.

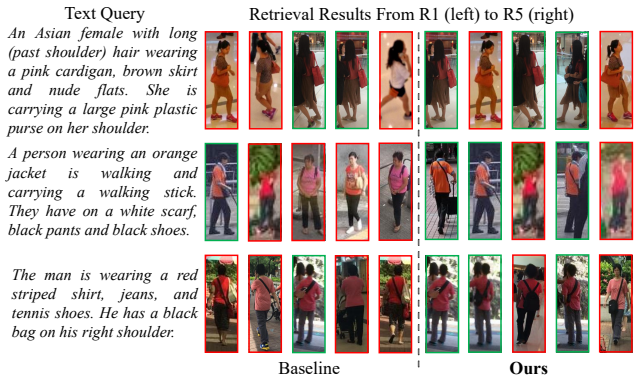


Figure 5: Visualization of Rank-5 results on the CUHK-PEDES dataset and each row of images is a group. The green and red bounding boxes represent the correct and false matches, respectively.

Conclusion

In this paper, we propose a framework named Diverse Person (DP) to achieve efficient and high-quality text-based person search data generation while avoiding privacy leakage issues. Specifically, we propose to edit both images and textual content of the original data using diffusion models and a Large Language Model (LLM). Furthermore, by supplementing the training set in a manner similar to mix-up data augmentation with reference attributes, DP can effectively enhance the models' ability to discriminate different pedestrian attributes and achieve stronger matching performance. Extensive experiments demonstrate the effectiveness and superiority of our approach compared to existing methods.

Ethical Statement

Text-based person search methods, like most technologies, have the potential to yield both societal benefits and negative consequences. For instance, by identifying target individuals, text-based person search can assist in apprehending suspects and counter-terrorism operations. However, the irresponsible implementation of this technology may invade personal privacy and its usage should be limited to public areas (*e.g.,* malls, airports, and parks). In this work, we achieve efficient and high-quality text-based person search data generation without involving privacy concerns and we strongly advocate for the development of ethical person search datasets. Additionally, the integration of ChatGPT in studies carries ethical implications with broad social ramifications. It enables inclusive communication but raises concerns about misinformation and biases. Ethical considerations demand transparency, bias mitigation, and ongoing evaluation to harness its benefits responsibly.

Acknowledgments

This work was supported by National Natural Science Fund of China (62076184, 61976158, 61976160, 62076182, 62276190), in part by Fundamental Research Funds for the Central Universities and State Key Laboratory of Integrated Services Networks (Xidian University), in part by Shanghai Innovation Action Project of Science and Technology (20511100700) and Shanghai Natural Science Foundation (22ZR1466700).

References

Chen, D.; Li, H.; Liu, X.; Shen, Y.; Shao, J.; Yuan, Z.; and Wang, X. 2018. Improving deep visual representation for person re-identification by global and local image-language association. In *Proceedings of the European conference on computer vision (ECCV)*, 54–70.

Chen, Y.; Huang, R.; Chang, H.; Tan, C.; Xue, T.; and Ma, B. 2021. Cross-modal knowledge adaptation for language-based person search. *IEEE Transactions on Image Processing*, 30: 4057–4069.

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*.

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2022. AXM-Net: Implicit cross-modal feature alignment for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4477–4485.

Han, X.; He, S.; Zhang, L.; and Xiang, T. 2021. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, R.; Sun, S.; Yu, X.; Xue, C.; Zhang, W.; Torr, P.; Bai, S.; and Qi, X. 2022. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hou, R.; Chang, H.; Ma, B.; Huang, R.; and Shan, S. 2021. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2014–2023.

Huang, S.-W.; Lin, C.-T.; Chen, S.-P.; Wu, Y.-Y.; Hsu, P.-H.; and Lai, S.-H. 2018. Auggan: Cross domain adaptation with gan-based data augmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 718–731.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1970–1979.

Linderman, G. C.; Rachh, M.; Hoskins, J. G.; Steinerberger, S.; and Kluger, Y. 2017. Efficient algorithms for t-distributed stochastic neighborhood embedding. *arXiv preprint arXiv:1712.09005*.

Liu, D.; Wu, L.; Hong, R.; Ge, Z.; Shen, J.; Boussaid, F.; and Bennamoun, M. 2022. Generative Metric Learning for Adversarially Robust Open-World Person Re-Identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*.

Niu, K.; Huang, Y.; Ouyang, W.; and Wang, L. 2020. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29: 5542–5556.

Niu, K.; Huang, Y.; and Wang, L. 2020. Textual dependency embedding for person search by language. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4032–4040.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021a. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5566–5574.

Shipard, J.; Wiliem, A.; Thanh, K. N.; Xiang, W.; and Fookes, C. 2023. Boosting zero-shot classification with synthetic data diversity via stable diffusion. *arXiv preprint arXiv:2302.03298*.

Shu, X.; Wen, W.; Wu, H.; Chen, K.; Song, Y.; Qiao, R.; Ren, B.; and Wang, X. 2022. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, 624–641. Springer.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Song, Z.; Zhao, C.; Hu, G.; and Miao, D. 2023. Learning Scene-Pedestrian Graph for End-to-End Person Search. *IEEE Transactions on Industrial Informatics*.

Suo, W.; Sun, M.; Niu, K.; Gao, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2022. A Simple and Robust Correlation Filtering Method for Text-Based Person Search. In *European Conference on Computer Vision*, 726–742. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Liang, X.; and Liao, S. 2022. 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*, 4900–4909.

Wang, Y.; Liao, S.; and Shao, L. 2020. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *Proceedings of the 28th ACM international conference on multimedia*, 3422–3430.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 2020. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 402–420. Springer.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022a. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5314–5322.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022b. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1984–1992.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 79–88.

Wu, Y.; Yan, Z.; Han, X.; Li, G.; Zou, C.; and Cui, S. 2021. LapsCore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1624–1633.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2023. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. *arXiv preprint arXiv:2305.10431*.

Xu, Y.; Ma, B.; Huang, R.; and Lin, L. 2014. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proc. 22th ACM Int. Conf. Multimedia*, 937–940.

Yan, R.; Tang, J.; Shu, X.; Li, Z.; and Tian, Q. 2018. Participation-contributed temporal dynamic model for group activity recognition. In *Proceedings of the 26th ACM international conference on Multimedia*, 1292–1300.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2023a. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*.

Yan, S.; Tang, H.; Zhang, L.; and Tang, J. 2023b. Image-specific information suppression and implicit local alignment for text-based person search. *IEEE Transactions on Neural Networks and Learning Systems*.

Yang, Y.; Wen, L.; Lyu, S.; and Li, S. 2017. Unsupervised learning of multi-level descriptors for person re-identification. In *Proc. AAAI Conf. Artif. Intell.*, volume 31.

Yao, R.; Gao, C.; Xia, S.; Zhao, J.; Zhou, Y.; and Hu, F. 2020. GAN-based person search via deep complementary classifier with center-constrained Triplet loss. *Pattern Recognition*, 104: 107350.

Zhang, T.; Xie, L.; Wei, L.; Zhuang, Z.; Zhang, Y.; Li, B.; and Tian, Q. 2021. Unrealperson: An adaptive pipeline towards costless person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11506–11515.

Zhang, Y.; and Lu, H. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, 686–701.

Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; and Shen, Y.-D. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2): 1–23.

Zhou, Q.; Zhang, K.-Y.; Yao, T.; Yi, R.; Sheng, K.; Ding, S.; and Ma, L. 2022. Generative domain adaptation for face anti-spoofing. In *European Conference on Computer Vision*, 335–356. Springer.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 209–217.

Zhu, L.; Yang, D.; Zhu, T.; Reda, F.; Chan, W.; Saharia, C.; Norouzi, M.; and Kemelmacher-Shlizerman, I. 2023. TryOnDiffusion: A Tale of Two UNets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4606–4615.