Diffusion-based Synthetic Data Generation for Visible-Infrared Person Re-Identification

Wenbo Dai[1], **Lijing Lu**[2,3*], **Zhihang Li**[3*]
[1]Nanjing Tech University, Nanjing, China
[2]Peking University, Beijing, China
[3]Chinese Academy of Sciences, Beijing, China
{a18151882515, lulijing1997, lizhihang.cas}@gmail.com

Abstract

The performance of models is intricately linked to the abundance of training data. In Visible-Infrared person ReIDentification (VI-ReID) tasks, collecting and annotating large-scale images of each individual under various cameras and modalities is tedious, time-expensive, costly and must comply with data protection laws, posing a severe challenge in meeting dataset requirements. Current research investigates the generation of synthetic data as an efficient and privacy-ensuring alternative to collecting real data in the field. However, a specific data synthesis technique tailored for VI-ReID models has yet to be explored. In this paper, we present a novel data generation framework, dubbed **Di**ffusion-based **V**I-ReID data **E**xpansion (**DiVE**), that automatically obtain massive RGB-IR paired images with identity preserving by decoupling identity and modality to improve the performance of VI-ReID models. Specifically, identity representation is acquired from a set of samples sharing the same ID, whereas the modality of images is learned by fine-tuning the Stable Diffusion (SD) on modality-specific data. DiVE extend the text-driven image synthesis to identity-preserving RGB-IR multimodal image synthesis. This approach significantly reduces data collection and annotation costs by directly incorporating synthetic data into ReID model training. Experiments have demonstrated that VI-ReID models trained on synthetic data produced by DiVE consistently exhibit notable enhancements. In particular, the state-of-the-art method, CAJ, trained with synthetic images, achieves an improvement of about 9% in mAP over the baseline on the LLCM dataset.

Introduction

Person Re-IDentification (ReID), aiming to match and identify individuals captured by non-overlapping cameras (Huang et al. 2023), has been widely used in numerous computer vision applications, including intelligent monitoring, public security, and persons analysis. Early efforts mainly focus on single-modality ReID tasks, where all the person images are typically collected by visible cameras under well-lit environments, ignoring the fact that visible cameras often fail to capture adequate information of one person under poor lighting conditions, restricting the applicability of the single-modality ReID for practical nighttime surveillance. To address this problem, researchers have introduced cross-modality ReID, known as Visible-Infrared Person ReIDentification (VI-ReID) (Ye et al. 2020a). The goal of VI-ReID is to match the night-time infrared person images captured by infrared (IR) cameras and visible person images captured by RGB cameras (Yang et al. 2023). With modern surveillance systems being able to automatically switch between visible and infrared modes during nighttime, VI-ReID has attracted more and more attention. Nevertheless, the distinct modality of infrared images presents substantial domain discrepancies, posing greater challenges for ReID.

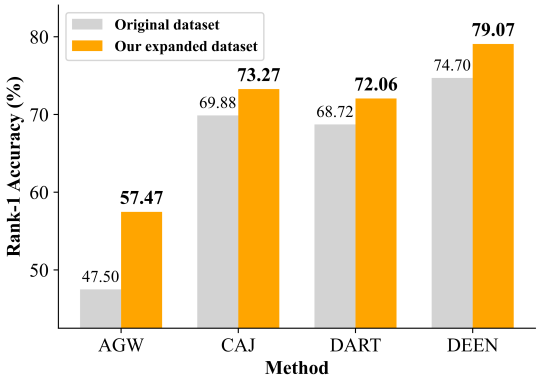


Figure 1: Performance comparison of different VI-ReID methods on the SYSU-MM01 dataset before and after using our proposed data expansion approach.

Several studies (Ye et al. 2020a; Fang, Yang, and Fu 2023) have made initial attempts of VI-ReID to learn a modality-invariant and discriminative feature space by designing more advanced neural network structures, improving objective functions, and adjusting training strategies. In this way, samples from the same person are closer across different modalities, while samples from different individuals are farther apart.Following above strategies, significant progress has been made in the field of VI-ReID (Ye et al. 2023; Zhang and Wang 2023; Ye et al. 2020a; Fang, Yang, and Fu 2023), but the performance of these methods is still limited by the scale of the dataset, leading to suboptimal accuracy. For example, the commonly used dataset SYSU-

*Corresponding authors.

Figure 2: The left image comes from the RGB-IR ReID dataset, while the right images come from the synthetic dataset generated by our method, DiVE. It can be seen that our method not only can synthesize images in the IR domain but also can maintain identity information. Especially, details such as backpacks, clothes, and hairstyles remain consistent with the RGB images. The generated images also exhibit different poses and scenes, enriching the diversity of datasets.

MM01 (Wu et al. 2020) only includes RGB and IR images of 491 identities from 6 cameras. Unfortunately, labeling the identity of each sample under different cameras and modalities is a labor-intensive and expensive process. Furthermore, there are situations where gathering images can pose challenges or may even be infeasible due to privacy and copyright constraints. Given the significant advancements in generative models (Rombach et al. 2022; Ramesh et al. 2021), an alternative approach involves leveraging synthetic data for training models (Zhang et al. 2021; Wu et al. 2023). However, synthesizing cross-modal person re-identification datasets is even more challenging. Due to the inherent domain gap between IR and RGB images, preserving identity consistency while accommodating modal differences is extremely challenging. For person re-identification, it is an open-set task with a large number of categories, each having very few samples. The few-shot nature of the data presents a challenge for training generative models.

To address the aforementioned issues, this paper proposes DiVE, an automatic procedure to generate a massive paired RGB-IR person images with different identities. Our method leverages cutting-edge zero-shot text-to-image models like Stable Diffusion, trained on extensive web-based image-text data. Given an arbitrary RGB image, DiVE identifies the corresponding IR textual description within the latent space of SD, enabling the creation of IR images with consistent identity. Fig.2 shows the results of our synthesized RGB-IR images. DiVE offers two key benefits tailored to address two distinct challenges. 1) *Identity preserving.* This paper utilizes Texture inversion to extract decoupled identity encodings from a group of images with the same ID. 2) *Modality Transformation.* Due to the rarity of IR images, it is not feasible to generate IR images directly through textual descriptions using SD. This paper proposes adapting Domain Customization, which fine-tunes SD with a small number of images to enable it to generate IR images. By decoupling identity and style, the proposed method can achieve RGB-to-IR image synthesis by modifying identity while maintaining consistency in identity. With the above advantages, DiVE can generate substantial RGB-IR paired images for any identities without human effort. These synthetic data can then be used for training any VI-ReID architectures. Our experiments demonstrate that DiVE can generate photorealistic and identity-consistent RGB-IR images. When training the VI-ReID model with synthetic data shown in Fig.1, there is a consistent and significant improvement in performance across different methods, such as the DEEN method achieving Rank-1 of 79.07%. Extensive ablation studies have investigated the impact of different modules of our method as well as dataset selection on the performance of synthetic data and ReID models. In summary, the contributions of this paper lie in the following aspects:

- We present a novel insight demonstrating the potential to automatically derive synthetic IR images while preserving identity by using a text-supervised pre-trained diffusion model. In light of this, we propose DiVE, an automatic procedure to generate massive RGB-IR paired images without human effort.
- The proposed DiVE is a framework that decouples identity and modality. Identity representation is learned from a group of samples with the same ID, while modality is endowed with the ability to generate synthetic IR images through fine-tuning the SD model on data within each modality. During inference, modifying identity and modality representations can synthesize new samples.
- Experiments demonstrate that VI-ReID models trained on syntheic dataset by DiVE have consistently shown a significant improvement. Extensive ablation studies have validated the effectiveness of each proposed module.

Related Work

Visible-Infrared Person Re-IDentification

The main challenges of VI-ReID lie in the significant discrepancies between two modalities and the intra-modality variations. Existing methods mainly focus on either learning modality-shared features or compensating for modality-specific information (Huang et al. 2023; Hao et al. 2019; Li et al. 2020). The former focus on extracting discriminative features shared across visible and infrared modalities. Key techniques include mining and aligning shared features from modality-specific features(DDAG (Ye et al. 2020a), MPANet(Wu et al. 2021), SAAI(Fang, Yang, and Fu 2023)) and optimizing the learning process with different loss functions or training strategies (BDTR(Ye et al. 2018), SFAnet(Chen et al. 2021), MCLNet(Hao et al. 2021), MAUM(Liu et al. 2022),DART(Yang et al. 2022)) The latter aim to generate missing modality-specific information to reduce the modality discrepancy. Representative works include utilizing generative models to transform modality from one to another(AlignGAN(Wang et al. 2019a), VI-Diff(Huang, Huang, and Wang 2023)), generating modality-specific features (D2RL(Wang et al. 2019b),FMCNet(Zhang et al. 2022))and mixing channel or part from data augmentation view(CAJ(Ye et al. 2023), DEEN(Zhang and Wang 2023), PartMix(Kim et al. 2023)).

Although the above methods have made significant progress, most of them primarily focus on designing more advanced network architectures or improving training strategies, few methods improve model performance from a data perspective. On one hand, the cost of manually annotating RGB-IR datasets is too high for large-scale implementation. On the other hand, generating high-fidelity RGB-IR human images synthetically is a very challenging task due to complex human body poses and significant domain gaps. In contrast, our method inherits the powerful generative capabilities of generative model pre-trained on large-scale data pairs, and explored a fine-tuning framework for decoupling identity and modality, achieving controllable synthesis of multimodal human body images.

Synthetic Dataset Generation

Synthetic datasets, generated via generative models, are increasingly used in various applications. This includes domain transfer(Zhu et al. 2017; Park et al. 2020), converting data from one domain to another, and data generation(Wang et al. 2024; Zhang et al. 2021; Li et al. 2022, 2023), aimed at enriching data diversity and supplementing rare samples. In the context of domain transfer, unpaired image-to-image translation methods have been widely explored, including GAN-based models like CycleGAN(Zhu et al. 2017) and CUT(Park et al. 2020), and diffusion-based methods such as ILVR(Choi et al. 2021) and EGSDE(Zhao et al. 2022). With the advancement of diffusion models, methods like DATUM(Benigmim et al. 2023), DOGE(Yinong Oliver Wang and la Torre 2024), and CycleGAN-Turbo(Parmar et al. 2024) employ pretrained text-to-image diffusion models for efficient domain transfer with limited examples. Existing image-to-image methods mainly focus on style transformation, while VI-ReID addresses a fine-grained recognition problem where identity preservation is crucial. Synthetic data is receiving increasing attention. Prior works (He et al. 2022) involved rendering images and labels using 3D game engines. With the development of generative models such as GANs(Zhang et al. 2021; Li et al. 2022; He et al. 2022; Wu et al. 2022) and Diffusion models (He et al. 2022; Wang et al. 2024), some work has started to utilize generative models to synthesize training data. Our approach is inspired by Diff-Mix (Wang et al. 2024), a method for fine-tuning text-to-image models for generating diverse samples interpolated between class. We take an innovative step forward by decoupling identity from modality information during the fine-tuning process, which enables us to freely combine identity and modality. As a result, we can generate any identity within the infrared modality.

Methodology

Preliminary and Problem Formulation

Text-to-Image Diffusion Model. Text-to-Image (T2I) diffusion models have emerged as a powerful tool for generating high-quality images conditioned on textual descriptions. Among these, the Stable Diffusion (SD) model is particularly notable. The SD model consists of a CLIP (Ramesh et al. 2021) text encoder Γ, and a U-Net (Ronneberger, Fischer, and Brox 2015) based conditional diffusion model ϵ_θ.

Given a text prompt Q, the text encoder Γ generates a conditioning vector $\Gamma(Q)$. During training, with a randomly sampled noise $\epsilon \sim \mathcal{N}(0, I)$ and the time step t, we can get a noised image or latent code $z_t = \alpha_t \mathbf{x} + \sigma_t \epsilon$, where $\mathbf{x}$ represents the input image, α_t and σ_t are the coefficients that control the noise schedule. The conditional diffusion model ϵ_θ is subsequently trained with the denoising objective:

$$E_{\mathbf{x},Q,\epsilon,t}\left[|\epsilon - \epsilon_\theta(z_t, t, \Gamma(Q))|_2^2\right] \tag{1}$$

Where ϵ_θ is trained to predict the noise condition on the noisy latent z_t, the text prompt Q, and the time step t. During inference, given a text prompt Q, we start from a noise latent $z_T \sim \mathcal{N}(0, I)$. The noise is gradually removed by iteratively predicting it for T steps using ϵ_θ. By the end of this process, we obtain a generated image that corresponds to the text prompt Q.

Formulation of VI-ReID. Formally, the training set of VI-ReID contains RGB and IR images, with each image x^k having identity labels y^k, where k denotes the modality and $k \in \{V, I\}$ (V for visible modality and I for infrared modality). The visible and infrared samples from the training set are denoted by:

$$\mathcal{V} = \{(x_n^V, y_n^V)\}_{n=1}^{N_V}, \quad \mathcal{I} = \{(x_n^I, y_n^I)\}_{n=1}^{N_I}$$

where N_V and N_I are the numbers of samples of each modality in the training set.

Let $\mathcal{P}_{\text{VI}} = \{p_n^{VI}\}_{n=1}^{N_{\text{VI}}}$ be the set of person identities in the training set, where N_{VI} and p represents the total number of identities and one person's identity, respectively. The identity labels $y^k \in \mathcal{P}_{\text{VI}}$.

The objective of Visible-Infrared Person ReIDentification is to match person identities across different modalities based on feature similarity. Therefore, it is essential to reduce the large intra-class variation between heterogeneous samples. Existing methods (Ye et al. 2020b, 2023) often address this by optimizing:

$$\mathcal{L} = \sum \ell\left(f\left(x_n^V\right), f\left(x_m^I\right), y_n^V, y_m^I\right) \tag{2}$$

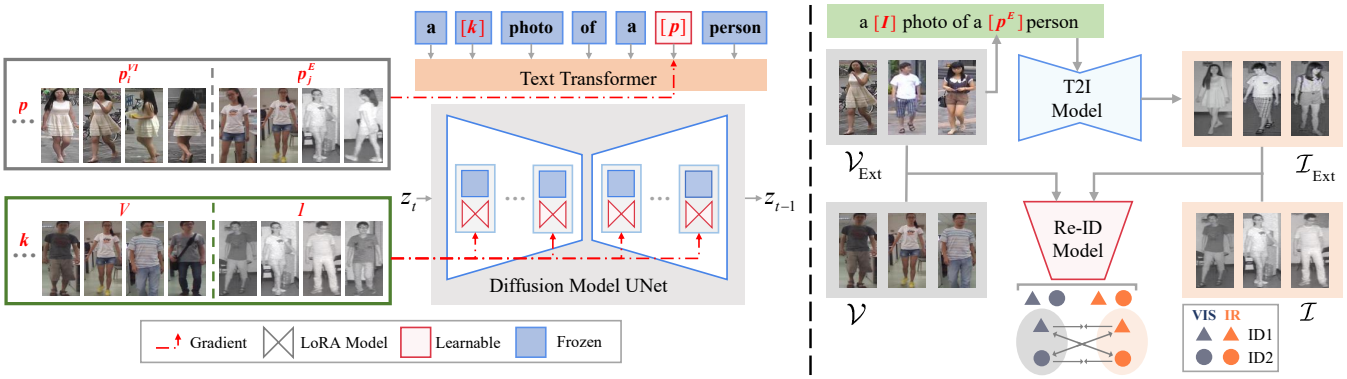
where f denotes the feature extractor, and $\ell(\cdot)$ represents a loss function such as identity loss or triplet loss.

Overview

To deal with the above problem, we propose a novel data generation framework, dubbed Diffusion-based VI-ReID data Expansion (DiVE), that automatically obtain massive RGB-IR paired images with identity preserving.

We aim to generate corresponding infrared counterparts for each identity within the external visible-based ReID dataset, thereby establishing a paired visible-infrared dataset to enrich the training data for the VI-ReID task.

Following the definition of VI-ReID mentioned above, we first define the external visible-based dataset as: $\mathcal{V}_{\text{Ext}} = \{(x_n^V, y_n^V)\}_{n=1}^{N_E}$, where N_E is the number of visible samples in the external dataset, and V denotes the visible modality. Let $\mathcal{P}_{\text{Ext}} = \{p_n^E\}_{n=1}^{N_{\text{Ext}}}$ denote the set of person identities in



(a) Learning implicit identity and modality textual descriptions

(b) Expanding dataset via generating infrared data

Figure 3: Illustration of our DiVE. (a): The training of the DiVE involves unpaired RGB-IR data. DiVE disentangles identity and modality representations to enrich the identity diversity of the generated images. (b): After training the generator, we leverage it to transfer a great deal of RGB images to IR images with identity preserved. These synthetic samples are used to train arbitrary VI-ReID approaches.

the external dataset, where N_{Ext} represents the total number of identities. The identity labels $y_n^V \in \mathcal{P}_{\text{Ext}}$.

Subsequently, synthetic infrared images are generated for each identity in the external visible-based dataset, yielding the set: $\mathcal{I}_{\text{Ext}} = \{(x_n^I, y_n^I)\}_{n=1}^{N_S}$, where each synthetic infrared image corresponds to an identity from the external visible-based dataset $\mathcal{P}_{\text{Ext}}$. Here, N_{S} denotes the number of synthetic infrared samples, and I represents the infrared modality. To enrich the existing VI-ReID dataset, we incorporate synthetic samples by merging $\mathcal{V}_{\text{Ext}}$ with the original visible dataset $\mathcal{V}$, and external visible data $\mathcal{I}_{\text{Ext}}$ with the original infrared dataset $\mathcal{I}$:

$$\mathcal{V}^* = \mathcal{V} \cup \mathcal{V}_{\text{Ext}}, \quad \mathcal{I}^* = \mathcal{I} \cup \mathcal{I}_{\text{Ext}}$$

The proposed DiVE

General Idea. To generate infrared images, we rely on the powerful SD model. As the images generated by SD depend on textual descriptions, and the textual description of each pedestrian image is unknown, we propose an inversion method to obtain the implicit textual description of each image. Given that cross-modality ReID images contain identity and modality information, we propose a unified mapping function $\mathcal{F}$, which can concurrently map images to the implicit textual embeddings of identity and modality.

Formally, given an identity $p \in \mathcal{P}_{\text{VI}} \cup \mathcal{P}_{\text{Ext}}$ and a modality $k \in \{V, I\}$, we have an image x_n^k with its label y_n^k, indicating identity $y_n^k = p$. We then have:

$$\mathcal{F} : x_n^k \rightarrow (\texttt{[}p\texttt{]}, \texttt{[}k\texttt{]}) \tag{3}$$

Here, $\texttt{[}p\texttt{]}$ is the implicit textual representation of the identity p, and $\texttt{[}k\texttt{]}$ is the implicit textual representation of the modality k.

In this way, we can obtain the modality invariant identity descriptions $\texttt{[}p^{\text{E}}\texttt{]}$ $(p^{\text{E}} \in \mathcal{P}_{\text{Ext}})$ by applying the function $\mathcal{F}$ to the external visible-based dataset $\mathcal{V}_{\text{Ext}}$.

Further, we could obtain the infrared modality description $\texttt{[}I\texttt{]}$ decoupled from the identity information by leveraging identity consistency. This is achieved by mapping images of the same identity p^{VI} from different modalities $(x_i^V \in \mathcal{V}, x_j^I \in \mathcal{I})$ to the same identity implicit textual description $\texttt{[}p^{\text{VI}}\texttt{]}$. Utilizing the function $\mathcal{F}$, this mapping can be represented as:

$$\mathcal{F} : (x_i^V, x_j^I) \rightarrow (\texttt{[}p^{\text{VI}}\texttt{]}, \texttt{[}V\texttt{]}, \texttt{[}I\texttt{]}) \tag{4}$$

Here, $\texttt{[}V\texttt{]}$ and $\texttt{[}I\texttt{]}$ are the modality descriptions for the visible and infrared images, respectively, learned by discriminating the discrepancy of the same identity across modalities.

After obtaining the implicit descriptions, we combine $\texttt{[}p^{\text{E}}\texttt{]}$ from the visible-based dataset $(p \in \mathcal{P}_{\text{Ext}})$ and $\texttt{[}I\texttt{]}$ from the IR modality to generate a text prompt. We use the function $\mathcal{T}(\cdot)$ to construct this text prompt. Subsequently, the SD model serves as the generator $\text{G}(\cdot)$ to transform these implicit descriptions into synthetic infrared images x^S:

$$x^S = \text{G}(\mathcal{T}(\texttt{[}p^{\text{E}}\texttt{]}, \texttt{[}I\texttt{]})) \tag{5}$$

Unified Mapping Function $\mathcal{F}$. T2I models not only excel at generating diverse images from textual descriptions but also possess the ability to invert specific styles or subjects of images back into the textual space through finetuning, a process known as personalization. Given a pre-trained T2I synthesis model and multiple images containing the target object, personalization involves optimizing a text prompt that represents the object, and optionally fine-tuning the network to better adapt to the target. After training, the optimized object text prompt can be combined with natural language descriptions to generate diverse outputs. In this paper, we leverage this personalization technique as our mapping function $\mathcal{F}$ to invert an image into a text prompt.

Specifically, we first employ the Textual Inversion (Gal et al. 2022) technique to extract identity descriptions by introducing a new learnable textual embedding, represented by the placeholder token $\texttt{[}p\texttt{]}$, for each identity p. The $\texttt{[}p\texttt{]}$ embedding resides in the textual embedding space and is shared among all images corresponding to the same identity,

regardless of their modality. During training, we optimize these embeddings to capture and store the modality-invariant identity information of each person p within their respective $[p]$ embedding. Then, we adopt a novel method to capture modality descriptions inspired by DreamBooth (Ruiz et al. 2023). Instead of introducing additional learnable tokens for each modality, we utilize fixed placeholder tokens $[k]$, such as $[\mathtt{V}]$ for visible and $[\mathtt{I}]$ for infrared. We finetune specific components of the model—namely, the Low-Rank Adaptation (LoRA (Hu et al. 2021)) modules associated with each modality—to encode modality-specific characteristics. LoRA proposes to finetune the network by introducing trainable low-rank residual matrices instead of updating the entire set of weights. During training, all images sharing the same modality, regardless of their identity, are assigned the same $[k]$ placeholder. This placeholder activates the corresponding LoRA branch, embedding modality information within these modules. After mapping, all images of an identity in a certain modality are converted into the text prompt "a $[k]$ photo of $[p]$ person." constructed by $\mathcal{T}(\cdot)$.

Finally, our loss function is defined as follows:

$$E_{\mathbf{x},p,\epsilon,t}\left[|\epsilon - \epsilon_\theta(z_t, t, \Gamma(\mathcal{T}([p], [k])))|_2^2\right] \quad (6)$$

where the optimizable parameters include the embeddings $\Gamma([p])$ for each id and the LoRA branches ϵ_θ.

Furthermore, in order to address hierarchical modality discrepancies, such as intra-camera and inter-camera variations (including viewpoint, illumination, and background differences), we refine the original $[V]$ and $[I]$ identifiers into more fine-grained categories: $[V_1]$, $[V_2]$, ..., $[V_m]$, $[I_1]$, $[I_2]$, ..., $[I_n]$, where m and n represent the number of camera views under visible and infrared camera individually, respectively. This detailed categorization allows us to describe the scene more precisely, promoting the decoupling of modality and identity information. We use the expanded IR modality identifiers $[I_1]$, $[I_2]$, ..., $[I_n]$ to generate IR images with clearer modality information across multiple views using a prompt of "a $[p^{\mathrm{E}}]$ photo of $[I_i]$ person." ($i = 1, \ldots, n$).

Experiments

Datasets and Evaluation Protocols

We evaluate our method on two VI-ReID datasets, namely SYSU-MM01(Wu et al. 2020), and LLCM(Zhang and Wang 2023), as well as two RGB person Re-IDentification datasets, including Market-1501(Zheng et al. 2015) and CUHK03-NP(Zhong et al. 2017; Li et al. 2014). Following common practices, we adopt the cumulative matching characteristics (CMC) and mean average precision (mAP) as evaluation metrics. Additionally, all the reported results are the average of 10 trails.

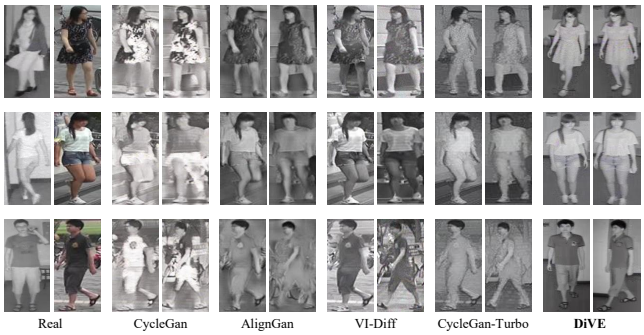


Figure 4: Visual comparison of synthetic infrared images. **Column 1:** Real IR images from SYSU-MM01 dataset and RGB images from Market-1501 dataset. **Columns 2-6:** Synthetic IR images generated by CycleGAN, AlignGAN, VI-Diff, CycleGAN-Turbo, and the proposed DiVE model, respectively.

Model	All-search		Indoor search	
	R1	mAP	R1	mAP
CycleGAN	70.98	67.93	78.40	82.34
Align-GAN	72.64	69.08	78.33	82.37
VI-Diff	73.24	69.44	79.81	83.23
CycleGAN-Turbo	73.12	69.76	80.37	83.65
Baseline (DEEN)	74.70	71.80	80.30	83.30
Our proposed DiVE	**79.07**	**74.96**	**82.98**	**85.90**

Table 1: Comparisons with GANs (CycleGAN, Align-GAN) and diffusion models (VI-Diff, CycleGAN-Turbo) on SYSU-MM01 dataset. The bold font denotes the best performance.

Implementation Details

Our method uses Stable Diffusion 1.5 as the base model, fine-tuning only the LoRA weights and textual embeddings. The rank of LoRA is set to 128, and each modality identifier is assigned a unique 8-character identifier (e.g. "b8zBXKoH"). During the training phase, all input images are resized to 512×256 pixels and augmented with horizontal flips to enhance model robustness. The learning rate is set to 5×10^{-5}. The batch size is configured to 16, and the total number of training steps is set to 400,000. For image generation, we utilize a timestep of 25 and adopt DPMsolver++ (Lu et al. 2022) as the sampling scheduler. We generate 18 infrared images for each modality identifier.

Comparison with State-of-the-Art Generative Models

We first compare DiVE with other prevalent methods for generating synthetic Infrared (IR) data. We specifically choose GAN-based methods (CycleGAN(Zhu et al. 2017), Align-GAN(Wang et al. 2019a)) and diffusion-based methods (CycleGAN-Turbo(Parmar et al. 2024), VI-Diff(Huang, Huang, and Wang 2023)). These methods are trained on

Methods	SYSU-MM01				LLCM			
	All search		Indoor search		VIS to IR		IR to VIS	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
AGW (Ye et al. 2020b)	47.50	47.65	54.17	62.97	49.13	55.80	**63.72**	47.21
+ DiVE	**57.47**	**56.18**	**63.72**	**70.54**	**51.60**	**58.53**	63.54	**50.21**
CAJ (Ye et al. 2023)	69.88	66.89	76.26	80.37	49.86	56.40	63.73	47.71
+ DiVE	**73.27**	**69.82**	**78.39**	**82.13**	**52.92**	**59.40**	**64.73**	**56.80**
DART (Yang et al. 2022)	68.72	66.29	72.52	78.17	52.97	59.28	65.33	51.13
+ DiVE	**72.06**	**68.83**	**76.38**	**81.12**	**54.89**	**61.24**	**65.47**	**59.59**
DEEN (Zhang and Wang 2023)	74.70	71.80	80.30	83.30	55.52	62.07	69.21	55.52
+ DiVE	**79.07**	**74.96**	**82.98**	**85.90**	**59.30**	**65.90**	**72.99**	**59.43**

Table 2: Comparisons with state-of-the-art methods on the SYSU-MM01 and LLCM datasets. The bold font denotes the best performance.

the RGB-IR SYSU-MM01 dataset. Then, we use these trained models to transfer RGB images from the Market-1501 dataset to the IR domain, expanding the dataset. Finally, we train and evaluate different models using these synthetic data on the DEEN VI-ReID model.

Qualitative Comparisons. Fig. 4 presents a visual comparison of the IR images generated by different methods. We observe that GAN-based methods (CycleGAN, Align-GAN) encounter a loss of semantic information and artifacts in the generated images. In contrast, diffusion models maintain superior semantic consistency owing to their stable training process. However, VI-Diff faces challenges in fully capturing IR modality characteristics. CycleGAN-Turbo, leveraging a fine-tuned pre-trained diffusion model, is capable of producing well-aligned images but is limited in scope. Our proposed DiVE model transcends these limitations, facilitating the generation of diverse, high-quality images compared to approaches focused solely on style transfer. This superiority can be attributed to a more fine-grained division of scenes within the same modality. Our generated images showcase a harmonious joint representation of human and environmental information in the IR modality.

Quantitative Analysis. Tab. 1 illustrates the ReID performance of model trained on data generated by different generative methods. Obviously, training with expanded data generated by either GAN or diffusion models leads to a decrease in model performance. GAN-based methods struggle to fit the distribution of real datasets accurately due to their unstable training and mode collapse. In contrast, diffusion models, characterized by a more stable training process, exhibit superior performance when compared to GAN-based approaches. Nonetheless, VI-Diff and CycleGan-Turbo methods merely learn the mapping between modalities without integrating the semantic consistency of the same ID across different modalities.

Our DiVE inherits the prior knowledge of the SD model trained on a large-scale dataset, ensuring the quality of synthesized images. The framework proposed in this paper for decoupling identity and mode ensures both the consistency of identity and adherence to the distribution of modes. Thus DiVE improving Rank-1 accuracy and mAP by 4.37% and 5.2%, respectively, in the All-search scenario.

Modality	View	SYSU		LLCM	
		R1	mAP	R1	mAP
-	-	74.70	71.80	55.52	62.07
IR	Single	75.59	71.89	56.21	62.75
RGB+IR	Single	76.89	72.98	55.78	62.93
RGB+IR	Multi	**79.07**	**74.96**	**59.30**	**65.90**

Table 3: Ablation study of modality and view selection.

Generalization of synthetic data

To demonstrate the effectiveness of our DiVE method, we augment data using DiVE on various datasets (SYSU-MM01, LLCM) and evaluate the performance using multiple VI-ReID models (AGW, CAJ, DART, DEEN). Table 2 presents the experimental results, and our synthetic data consistently improves model performance across both SYSU-MM01 and LLCM. CAJ even shows an improvement of up to 9% mAP on the LLCM. This indicates that our synthetic data closely approximates the true IR domain distribution, aiding the models in learning a more generalized discriminative latent space.

From model perspective, our synthetic data improves the performance of both modality-shared feature learning-based methods (AGW, DART) and modality-specific information compensation-based methods (CAJ, DEEN). For modality-shared feature learning-based methods, which rely on comparing the same ID across different modalities to learn modality-invariant features, our augmented data strengthens and diversifies the modality-invariant features by providing more paired images for each ID. On the other hand, modality-specific information compensation-based methods require a large amount of data to compensate for the missing modality information, and our augmented data enhances the diversity of samples.

Ablation Studies

Impact of modality and view selection. Here we study the impact of different data selections on DiVE. In addi-

Dataset selection	SYSU		LLCM	
	R1	mAP	R1	mAP
-	74.70	71.80	55.52	62.07
+CUHK-NP	**77.50**	**73.48**	**58.87**	**65.47**
+Market1501	**79.07**	**74.96**	**59.30**	**65.90**

Table 4: Ablation study of using different visible-based dataset.

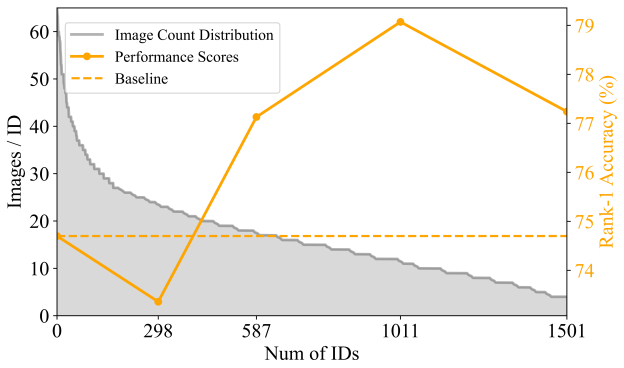


Figure 5: Performance under different number of Augmented IDs. Different colors represent different identities.

tion to RGB and IR affecting model training, images from different views within the same modality also exhibit different distributions. We consider three training sets: 1) only IR images; 2) RGB+IR images under single view ; 3) multi-view of RGB+IR images. When only using IR modality, the data generated by DiVE is still beneficial for ReID models, surpassing the baseline. The model's performance is further improved by adding the RGB modality, indicating enhanced learning of modality descriptions. It is worth noting that the most significant performance gains by considering multiple views within each modality, as shown in Tab. 3, resulting in Rank-1 and mAP scores of 79.07% and 74.96% on SYSU-MM01, and 60.30% and 66.90% on LLCM. In this way, each view serving as a more nuanced modality allows the model to learn more precise modality information.

Influence of Different Visible modality ReID Datasets. We investigate the impact of utilizing different RGB datasets (Market-1501, CUHK-NP) to expand the training data and evaluate their performance on the SYSU-MM01 and LLCM dataset. Tab. 4 shows that incorporating Market-1501 and CUHK-NP significantly outperforms the baseline on both bath SYSU and LLCM datasets. The results indicate our synthetic data is effective and versatile, as it can benefit from various visible-based datasets to enhance VI-ReID performance.

Number of Augmented IDs. We explore the effect of the number of IDs selected from the RGB dataset for expansion. Figure 5 illustrates the impact of the augmented ID number on VI-ReID performance. The grey line represents the distribution of image counts per ID, while the red line shows

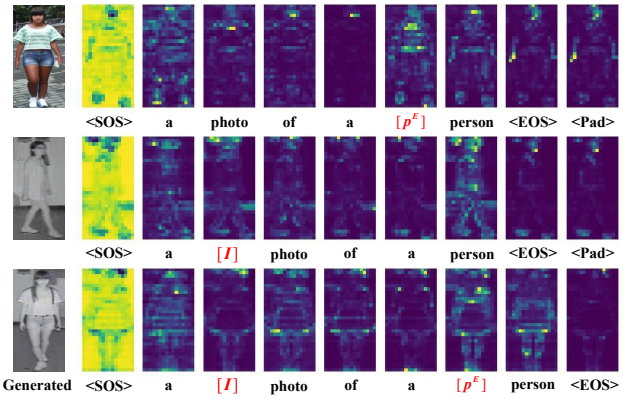


Figure 6: Visualizations of attention maps corresponding to different prompt

the rank-1 accuracy of model at different numbers of augmented IDs. The results suggest that increasing the number of augmented IDs initially leads to significant performance gains, with accuracy improving from the baseline of 74.70% to a peak of 79.07% when 1011 IDs are used. However, continuing to increase the number of IDs will harm the model's performance. The underlying reason maybe that due to the long-tail distribution of the dataset, extracting ID features from a small number of images is challenging. Therefore, finding an optimal balance is important for data generation.

Effectiveness of textual control. We visualized the cross attention map corresponding to each token to analyze their impact on the synthesized images. Figure 6 displays three types of prompts: only identity, only modality, and containing both identity and modality simultaneously. We observe that modality token emphasizes global features, including both person and background regions, while the ID token focuses more on the person's area, emphasizing person-specific information. Comparing synthetic images, altering $[k]$ changes the modality without affecting identity, and modifying $[p]$ changes the identity while preserving modality.

Conclusion

This paper introduces a novel perspective whereby large-scale RGB-IR images with consistent identities are automatically generated using a text-driven diffusion model, enhancing the performance of person Re-IDentification (ReID) models. To accomplish this objective, we present Diffusion-based VI-ReID data Expansion (DiVE), an innovative data generation framework that automatically produces massive RGB-IR paired images while preserving identities without human intervention, achieved by decoupling identity and modality. Extensive experiments demonstrate that VI-ReID models trained on synthetic data generated by the DiVE framework show superior performance compared to those trained on real data. Additionally, we anticipate that DiVE can bring fresh insights and inspiration for bridging generative data and person ReID in the community.

References

Benigmim, Y.; Roy, S.; Essid, S.; Kalogeiton, V.; and Lathuilière, S. 2023. One-shot Unsupervised Domain Adaptation with Personalized Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 698–708.

Chen, Y.; Wan, L.; Li, Z.; Jing, Q.; and Sun, Z. 2021. Neural feature search for rgb-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 587–597.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*.

Fang, X.; Yang, Y.; and Fu, Y. 2023. Visible-Infrared Person Re-Identification via Semantic Alignment and Affinity Inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11270–11279.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Hao, X.; Zhao, S.; Ye, M.; and Shen, J. 2021. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International conference on computer vision*, 16403–16412.

Hao, Y.; Wang, N.; Li, J.; and Gao, X. 2019. HSME: Hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8385–8392.

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*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, H.; Huang, Y.; and Wang, L. 2023. VI-Diff: Unpaired Visible-Infrared Translation Diffusion Model for Single Modality Labeled Visible-Infrared Person Re-identification. arXiv:2310.04122.

Huang, N.; Liu, J.; Miao, Y.; Zhang, Q.; and Han, J. 2023. Deep learning for visible-infrared cross-modality person re-identification: A comprehensive review. *Information Fusion*, 91: 396–411.

Kim, M.; Kim, S.; Park, J.; Park, S.; and Sohn, K. 2023. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18621–18632.

Li, D.; Ling, H.; Kim, S. W.; Kreis, K.; Fidler, S.; and Torralba, A. 2022. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21330–21340.

Li, D.; Wei, X.; Hong, X.; and Gong, Y. 2020. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 4610–4617.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *CVPR*.

Li, Z.; Zhou, Q.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Guiding text-to-image diffusion model towards grounded generation. *arXiv preprint arXiv:2301.05221*, 3(6): 7.

Liu, J.; Sun, Y.; Zhu, F.; Pei, H.; Yang, Y.; and Li, W. 2022. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19366–19375.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. *arXiv preprint arXiv:2206.00927*.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive Learning for Unpaired Image-to-Image Translation. In *European Conference on Computer Vision*.

Parmar, G.; Park, T.; Narasimhan, S.; and Zhu, J.-Y. 2024. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

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.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; and Hou, Z. 2019a. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3623–3632.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2019b. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 618–626.

Wang, Z.; Wei, L.; Wang, T.; Chen, H.; Hao, Y.; Wang, X.; He, X.; and Tian, Q. 2024. Enhance Image Classification

via Inter-Class Image Mixup with Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17223–17233.

Wu, A.; Zheng, W.-S.; Gong, S.; and Lai, J. 2020. RGB-IR person re-identification by cross-modality similarity preservation. *International journal of computer vision*, 128(6): 1765–1785.

Wu, Q.; Dai, P.; Chen, J.; Lin, C.-W.; Wu, Y.; Huang, F.; Zhong, B.; and Ji, R. 2021. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4330–4339.

Wu, W.; Zhao, Y.; Shou, M. Z.; Zhou, H.; and Shen, C. 2023. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1206–1217.

Wu, Z.; Wang, L.; Wang, W.; Shi, T.; Chen, C.; Hao, A.; and Li, S. 2022. Synthetic data supervised salient object detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5557–5565.

Yang, B.; Chen, J.; Ma, X.; and Ye, M. 2023. Translation, association and augmentation: Learning cross-modality re-identification from single-modality annotation. *IEEE Transactions on Image Processing*.

Yang, M.; Huang, Z.; Hu, P.; Li, T.; Lv, J.; and Peng, X. 2022. Learning With Twin Noisy Labels for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14308–14317.

Ye, M.; Shen, J.; Crandall, D. J.; Shao, L.; and Luo, J. 2020a. Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-Identification. In *European Conference on Computer Vision (ECCV)*.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. H. 2020b. Deep Learning for Person Re-identification: A Survey and Outlook. *arXiv preprint arXiv:2001.04193*.

Ye, M.; Wang, Z.; Lan, X.; and Yuen, P. C. 2018. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, volume 1, 2.

Ye, M.; Wu, Z.; Chen, C.; and Du, B. 2023. Channel Augmentation for Visible-Infrared Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yinong Oliver Wang, C. H. W., Younjoon Chung; and la Torre, F. D. 2024. Domain Gap Embeddings for Generative Dataset Augmentation. In *CVPR*.

Zhang, Q.; Lai, C.; Liu, J.; Huang, N.; and Han, J. 2022. FMCNet: Feature-Level Modality Compensation for Visible-Infrared Person Re-Identification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7339–7348.

Zhang, Y.; Ling, H.; Gao, J.; Yin, K.; Lafleche, J.-F.; Barriuso, A.; Torralba, A.; and Fidler, S. 2021. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10145–10155.

Zhang, Y.; and Wang, H. 2023. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2153–2162.

Zhao, M.; Bao, F.; Li, C.; and Zhu, J. 2022. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 1116–1124.

Zhong, Z.; Zheng, L.; Cao, D.; and Li, S. 2017. Re-ranking Person Re-identification with k-reciprocal Encoding. In *CVPR*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.