Reffusion: Self-Boosted Person Re-Identification via Pose-Guided Person Generation

Yong Zhao[1,2,3], Yali Li[1,2,3] and Shengjin Wang[1,2,3]*
[1]*Department of Electronic Engineering, Tsinghua University, China*
[2]*Beijing National Research Center for Information Science and Technology (BNRist), China*
[3]*National Engineering Research Center of Dangerous Articles and Explosives Detection Technologies, China*
y-zhao23@mails.tsinghua.edu.cn, {liyali13,wgsgj}@tsinghua.edu.cn

***Abstract*—Person re-identification (ReID) focuses on learning identity-discriminative and pose-invariant features. To enrich pose diversity of training data, person synthesis techniques are employed to generate images with novel poses. However, existing diffusion-based methods struggle to maintain identity consistency and decouple the person from background interference, often producing low-quality images which hinder ReID training. To address these limitations, we present Reffusion, a diffusion-based framework with an Online Filtering and Training Policy (OFTP). Reffusion concentrates on generating high-fidelity person images, while OFTP aims to train ReID model with generated persons. Specifically, Reffusion integrates Orthogonal Memory Module to capture decorrelated human patterns and Adaptive Masking Training to mitigate background interference. During inference of Reffusion, Instance Gradient Guidance is utilized to enhance identity consistency. When training ReID with these generated persons, OFTP adaptively filters out low-quality samples and retains informative ones for effective post-training. Experiments demonstrate that Reffusion yields an 8.008 FID on Market1501, while OFTP achieves 1.6% and 6.4% improvements in mAP over the strong baseline BoT on MSMT17 and CUHK03, respectively, significantly outperforming state-of-the-art methods.**

***Index Terms*—Person Re-identification, Conditional Diffusion, Person Synthesis, On-Policy Training**

I. Introduction

Person re-identification (ReID) [1] aims to recognize the same person across variations in camera, background, viewpoint, and other factors. Most existing works [2]–[5] focus on learning identity-discriminative features through multi-granularity representation alignment. While these approaches have achieved remarkable progress, their performance remains sub-optimal due to limited annotated training data. Furthermore, collecting a large-scale labeled person dataset is not only labor-intensive but also raises privacy concerns, limiting the scalability and applicability of ReID in real-world scenarios.

An intuitive approach to circumventing insufficient data is to effectively leverage synthetic person images. Pose-Guided Person Image Synthesis (PGPIS) aims to generate novel persons conditioned on a reference image and a target pose, while maintaining identity and texture consistency. By augmenting training data with new poses, PGPIS is expected to enhance ReID performance. Existing synthesis methods can be broadly categorized into into GAN-based [6] and diffusion-based [7] models. GAN-based approaches such as PN-GAN [8], DPTN [9] and NTED [10] often suffer from blurry and low-quality synthesis, primarily due to the instability of adversarial training and one-step generation. In contrast, diffusion [11] models have surpassed GANs in generating high-fidelity images through hierarchical VAE modeling. Consequently, several diffusion-based methods [12]–[16] have been proposed for PGPIS and achieved remarkable progress, but challenges remain in detail fidelity and identity preservation. For example, PIDM [12] and Identity Diffuser [16] employ simple conditioning or coarse identity guidance, restricting fine-grained synthesis. Methods such as CFLD [13], PCDMs [14] and IMAGPose [15] are trained in VAE latent space, which often causes detail loss for low-resolution pedestrian images (Figure 1). Moreover, some diffusion-based methods have overlooked the impact of background on person synthesis, resulting in distorted and chaotic persons (as illustrated in row 1 of Figure 1a). When utilizing generated images for ReID training, Diff-ID [17] directly adds generated images into original training datasets, whereas PIDM [12], PCDMs [14] and Identity Diffuser [16] fine-tune backbone (trained on real datasets) with generated images. However, these approaches haven't filtered low-quality or problematic images, in which identity preservation and pose transfer fail. The inclusion of such low-quality images may degrade ReID performance.

(a) Synthesized examples 1. (b) Synthesized examples 2.

Fig. 1. Visual comparison of different methods on person synthesis. CFLD suffers from detail loss due to VAE downsampling, while PIDM exhibits artifacts caused by missing 2D joints and background interference.

To address the issues of existing diffusion-based methods, we propose **Re**ID Di**ffusion** (Reffusion) along with an **O**nline

*Corresponding author.
This work was supported by the National Natural Science Foundation of China under Grant Nos. 61701277 and 61771288.

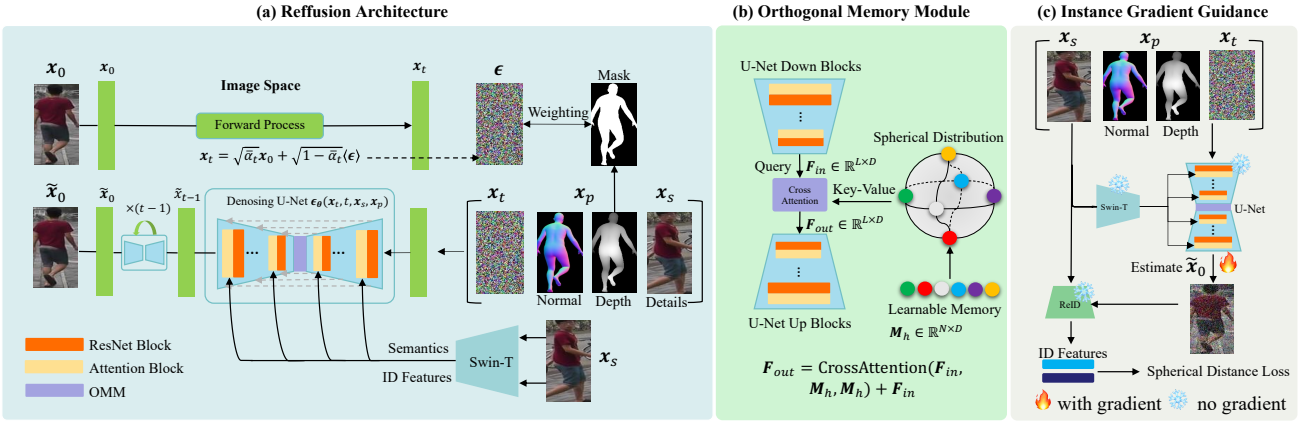

Fig. 2. (a) The overall architecture of proposed Reffusion, combined with Adaptive Masking Training (AMT); (b) Orthogonal Memory Module (OMM) is inserted in the middle of UNet, which enables Reffusion capture orthogonal human patterns by applying spherical distribution regularization; (c) during DDIM sampling, Instance Gradient Guidance (IGG) further improves identity consistency through gradient guidance from a frozen lightweight ReID network.

Filtering and **T**raining **P**olicy (OFTP). Reffusion generates high-fidelity persons with superior identity preservation and pose transfer. Then, OFTP adaptively filters out low-quality generated images for robust ReID post-training. Our contributions can be summarized as follows:

- We propose Reffusion, that generates high-quality person images with enhanced identity preservation and reduced background interference, by incorporating three key modules: Orthogonal Memory Module, Adaptive Masking Training, and Instance Gradient Guidance.
- To effectively leverage generated images for robust ReID post-training, we adopt an Online Filtering and Training Policy that selects high-quality and hard positive samples, while adaptively filtering out low-quality samples.
- Extensive experiments show that Reffusion achieves an 8.008 FID on Market1501, while OFTP brings 1.6% and 6.4% gains in mAP over the strong baseline BoT on MSMT17 and CUHK03, respectively, achieving state-of-the-art performance on both image synthesis and ReID.

II. Methodology

A. Architecture and Overview

Our method is built upon the UNet-based Denoising Diffusion Probabilistic Model (DDPM) [7], which consists of a **forward diffusion process** and a **reverse denoising process**. During the forward process, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0,1)$ is gradually added to the input image $\boldsymbol{x}_0$, according to a timestep t and a predefined noise schedule $\overline{\alpha}_t$:

$$\boldsymbol{x}_t = \sqrt{\overline{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, t \sim U[1,1000]. \tag{1}$$

In the denoising phase, UNet model $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ learns to reconstruct $\boldsymbol{x}_0$ from a noisy sample $\boldsymbol{x}_T$ conditioned on c. After applying reparameterization trick, the final training objective is reduced to a noise prediction loss:

$$\mathcal{L}_{mse} = \mathbb{E}_{\boldsymbol{x}_0,c,\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t,c)\|_2^2\right], \tag{2}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ predicts added noise $\boldsymbol{\epsilon}$ at each timestep t.

Figure 2 illustrates the overall architecture of our proposed Reffusion. Reffusion is trained in image space without VAE downsampling and predicts target image $\boldsymbol{x}_0$ conditioned on a reference image $\boldsymbol{x}_s$ and target pose $\boldsymbol{x}_p$, where $\boldsymbol{x}_0$ and $\boldsymbol{x}_s$ share the same identity. $\boldsymbol{x}_s$ is passed through a swin transformer [18] to extract multi-scale and semantically rich global features. The UNet captures these identity-related features via cross-attention. Also, $\boldsymbol{x}_s$ is concatenated with the target pose $\boldsymbol{x}_p$ and noisy samples $\boldsymbol{x}_t$ as input to preserve fine-grained local details such as edges and textures. In contrast to prior works [12]–[14] that rely solely on 2D sparse human skeletons as pose, Reffusion takes both depth map and surface normal map $\boldsymbol{x}_p$ detected from 3D SMPL [19] as pose prompt, which circumvents missing joints inherent in 2D and offers more accurate viewpoint cues. Moreover, Reffusion enhances the quality of generated person images through three core components: Orthogonal Memory Module, Adaptive Masking Training, and Instance Gradient Guidance. Ultimately, an Online Filtering and Training Policy (OFTP) adaptively removes low-quality generated samples, enabling more robust ReID training. The details will be elaborated in the following sections.

B. Orthogonal Memory Module

According to spherical coding theory [20], points (or feature vectors) on the unit hypersphere should be arranged to maximize mutual separation, which promotes diversity and reducing redundancy. Motivated by this principle, we introduce an Orthogonal Memory Module (OMM) in the middle of Reffusion. As illustrated in Figure 2 (b), OMM contains N learnable embeddings $\boldsymbol{M}_h$, which interact with hidden states of UNet via a cross attention mechanism.

During training, we employ a uniformity regularization term [21] that encourages these embeddings to be mutually decorrelated and nearly orthogonal, effectively enabling the memory to capture diverse human patterns. Specifically, the memory embeddings are first ℓ_2-normalized and then regularized as:

$$\mathcal{L}_{\text{reg}} = \log \mathbb{E}_{i\neq j}\left[\exp\left(-2\left\|\boldsymbol{m}_h^i - \boldsymbol{m}_h^j\right\|_2^2\right)\right]. \tag{3}$$

where $\boldsymbol{m}_h^i$ denotes the i^{th} embedding in OMM, and the expectation is taken over all distinct pairs $\{(\boldsymbol{m}_h^i, \boldsymbol{m}_h^j) \mid 1 \leq i \neq j \leq N\}$ within the memory.

By explicitly storing these disentangled human patterns, the diffusion model can query the memory to retrieve relevant structured or semantic information during the generation process. This mechanism ensures that each generated sample maintains realistic body shapes and textures, producing more consistent human images.

C. Adaptive Masking Training

In real-world scenarios, pedestrians are often accompanied by background clutter. To disentangle the synthesis of the person from such interference, we introduce Adaptive Masking Training (AMT) that re-weights the noise prediction, thereby encouraging the model to concentrate more on the foreground person. Specifically, as demonstrated in Figure 2 (a), a binary mask $\boldsymbol{A}_p$ can be derived from pose $\boldsymbol{x}_p$ through a binarization operation. $\boldsymbol{A}_p$ is then rescaled by the inverse mask area ratio and shifted by one, yielding $\boldsymbol{A} = r \cdot \boldsymbol{A}_p + 1$ where $r = \text{Area(image)}/\text{Area(mask)}$. This design adaptively assigns larger weights to smaller persons in the image, thereby emphasizing their contribution during training.

By integrating $\boldsymbol{A}$ with Equations (2) and (3), the overall training loss is finally formulated as:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_p, \boldsymbol{x}_s, \boldsymbol{\epsilon}, t} \left[\boldsymbol{A} \odot \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{x}_p, \boldsymbol{x}_s)\|_2^2 \right] + \lambda \mathcal{L}_{\text{reg}}. \tag{4}$$

During training, classifier-free guidance (CFG) [22] is adopted to strengthen source image guidance. $\boldsymbol{x}_s$ is randomly dropped with a probability of $\eta\%$, which encourages generated $\boldsymbol{x}_0$ to more closely resemble $\boldsymbol{x}_s$ in appearance. Thus, the sampling equation can be calculated by ($\boldsymbol{x}_p$ is omitted for simplicity):

$$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}} = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \varnothing) + w_{\text{cfg}}(\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{x}_s) - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \varnothing)). \tag{5}$$

D. Instance Gradient Guidance

In addition to CFG, we incorporate an external pre-trained ReID network to further enhance identity consistency between the generated person image $\boldsymbol{x}_0$ and the reference instance $\boldsymbol{x}_s$. We refer to this strategy as Instance Gradient Guidance (IGG). Inspired by [11], a score function $\nabla_{\boldsymbol{x}_0} p(\boldsymbol{x}_0)$ is used to guide the denoising process, as demonstrated in Figure 2 (c). The estimated $\tilde{\boldsymbol{x}}_0$ is firstly derived from $\boldsymbol{x}_t$ by:

$$\tilde{\boldsymbol{x}}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\overline{\alpha}_t}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{x}_p, \boldsymbol{x}_s)}{\sqrt{\overline{\alpha}_t}}. \tag{6}$$

Then, both the reference image $\boldsymbol{x}_s$ and estimated $\tilde{\boldsymbol{x}}_0$ are fed into a pre-trained ReID network to compute spherical distance:

$$\mathcal{L}_{\text{sph}}(f_{\boldsymbol{x}_s}, f_{\tilde{\boldsymbol{x}}_0}) = 2 \cdot \left(\arcsin\left(\frac{\|f_{\boldsymbol{x}_s} - f_{\tilde{\boldsymbol{x}}_0}\|_2}{2} \right) \right)^2, \tag{7}$$

where $f_{\boldsymbol{x}_s}$ and $f_{\tilde{\boldsymbol{x}}_0}$ are identity features extracted by ReID network and then l_2-normalized. Finally, new $\tilde{\boldsymbol{x}}_0$ is updated via gradient descent on $\mathcal{L}_{\text{sph}}(f_{\boldsymbol{x}_s}, f_{\tilde{\boldsymbol{x}}_0})$.

$$\text{new}(\tilde{\boldsymbol{x}}_0) = \tilde{\boldsymbol{x}}_0 - w_{\text{igg}}\sqrt{1-\overline{\alpha}_t} \cdot \nabla_{\tilde{\boldsymbol{x}}_0} \mathcal{L}_{\text{sph}}(f_{\boldsymbol{x}_s}, f_{\tilde{\boldsymbol{x}}_0}), \tag{8}$$

Algorithm 1: On-Policy Training with Generated Samples

Input: initialized ReID model π_θ; real dataset $\mathcal{D}_r$; generated dataset $\mathcal{D}_g$; selected dataset $\mathcal{D}_f \leftarrow \varnothing$; filtering threshold τ; k-nearest

Train π_θ on $\mathcal{D}_r$ for a number of epochs
Extract representations f_i^r for each sample in $\mathcal{D}_r$ by π_θ
Compute centroid features f_c^r for each ID based on f_i^r
for *sample in* $\mathcal{D}_g$ **do**
    $f_i^g \leftarrow \pi_\theta(sample)$
    Find k-nearest neighbors $\{f_{c_i}^r\}_{c_i=1}^k$ in f_c^r for f_i^g
    if #ID(f_i^g) in #ID($\{f_{c_i}^r\}_{c_i=1}^k$) & similarity $> \tau$ **then**
        $\mathcal{D}_f$.add($sample$)
Continue training π_θ on $\mathcal{D}_r$ and $\mathcal{D}_f$ until end

where w_{igg} denotes guidance scale and $\tilde{\boldsymbol{x}}_0$ is gradient-enabled. The next-timestep prediction $\tilde{\boldsymbol{x}}_{t-1}$ can be obtained by:

$$\tilde{\boldsymbol{x}}_{t-1} = \text{scheduler}(\text{new}(\tilde{\boldsymbol{x}}_0), \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}, t), \tag{9}$$

where scheduler can be either DDPM [7] or DDIM [23] sampler. Notably, IGG module introduces about 10% inference time while requiring minimal additional memory, since back-propagation is restricted to the ResNet-50 [24] ReID network.

E. On-Policy Post-Training with Generated Samples

Previous works [12], [14]–[16] that incorporate generated samples into ReID post-training typically follow an off-policy paradigm, where these samples are directly used for fine-tuning without any filtering. Such strategy is sub-optimal, since generated samples inevitably contain artifacts. To address this issue, we propose an Online Filtering and Training Policy (OFTP), which adaptively selects samples during training to improve robustness. As demonstrated in Algorithm 1, ReID model π_θ is initially trained on real dataset $\mathcal{D}_r$. We then compute the centroid features by averaging f_i^r within each ID cluster of $\mathcal{D}_r$ and extract the generated feature f_i^g for each sample in generated dataset $\mathcal{D}_g$. To effectively mine hard generated samples, the k-nearest neighbors $\{f_{c_i}^r\}_{c_i=1}^k$ of f_i^g are retrieved in the centroid feature space using cosine distance metric. A generated sample is retained if the identity of f_i^g matches at least one of the identities in $\{f_{c_i}^r\}_{c_i=1}^k$ and its corresponding cosine similarity exceeds threshold τ. Finally, the selected hard and reliable cases are combined with $\mathcal{D}_r$ for continual training, while bad unreliable cases are discarded. In this way, the overall distribution of the training dataset is preserved without being distorted by $\mathcal{D}_f$.

III. Experiments

A. Experiments Details

Datasets. Experiments are conducted on three public ReID datasets: Market1501 [25], MSMT17 [26] and CUHK03 [27]. Market1501 contains 12,936 training images of 751 identities captured by 6 cameras; MSMT17 has 32,621 training images of 1,041 identities collected from 15 cameras; and CUHK03

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON PERSON SYNTHESIS QUALITY.

Datasets	Methods	FID↓	SSIM↑
Market1501 (128 × 64)	PIDM [12]	14.451	0.3054
	Identity Diffuser [16]	13.290	–
	PCDMs [14]	13.897	0.3169
	CFLD [13]	11.972	0.3173
	IMAGPose [15]	12.659	0.3282
	Reffusion (Ours)	**8.008**	**0.3293**

includes 7,368 images of 767 identities taken from 2 cameras. Paired samples $(\boldsymbol{x}_s, \boldsymbol{x}_0)$ for each identity are densely sampled from corresponding ReID training dataset, while the target pose $\boldsymbol{x}_p$ is extracted by off-the-shelf 4D-Humans [19].

Implementation details. For person synthesis, all experiments are conducted on 4 Tesla V100 GPUs. The configuration details are summarized as follows:

- We adopt a lightweight UNet from [11] for diffusion and Swin-Tiny [18] as reference image encoder. The UNet contains approximately 170M parameters, which is about one quarter of stable diffusion v1.4[1].
- For Market1501, the image size is 128×64, while for MSMT17 and CUHK03, the resolution is set to 256×128.
- Reffusion is trained for 100 epochs with batch size 32 per device. Adam optimizer is used with cosine learning rate decay, starting from 2e-5. λ in Equation (4) is 0.1.
- The number of embeddings in OMM is fixed to 64. A pre-trained BoT Baseline-S model [3] is employed to extract identity features in Equation (7).
- During sampling, we use a DDIM scheduler with 50 steps and set the CFG and IGG scales to 2 and 20, respectively.

For person re-identification, 2x synthetic images are firstly produced with target poses randomly sampled from training set, after which OFTP is applied for ReID post-training.

Evaluation metrics. Fréchet Inception Distance (FID) [28] and Structural Similarity Index Measure (SSIM) [29] are used to assess the realism of generated images, where FID measures the distributional discrepancy between real and generated images, and SSIM assesses pixel-level structural consistency. Additionally, mean Average Precision (mAP) is utilized to evaluate ReID performance.

B. Quantitative and Visual Results on Person Synthesis

Table I demonstrates that Reffusion achieves state-of-the-art person synthesis quality on Market1501, outperforming CFLD [13] with a 3.964 reduction in FID score and achieving a higher SSIM metric to IMAGPose [15]. As shown in Figure 3, Reffusion preserves more fine-grained local details and accurately transfers the source image to the target pose while maintaining strong identity consistency. In contrast, PIDM and CFLD tend to produce blurry regions (e.g., faces and clothing) and distorted legs due to missing 2D human joints.

[1]https://huggingface.co/CompVis/stable-diffusion-v1-4

Fig. 3. Visual comparison with state-of-the-art methods on person synthesis quality. The green boxes indicate the shortcomings of prior methods, while the red boxes illustrate the advantages of Reffusion.

C. Performance on Person Re-identification

As illustrated in Table II, we evaluate OFTP with images generated by Reffusion across diverse ReID methods and datasets. OFTP consistently improves ReID performance on different backbones, from CNNs such as BoT [3] and LUP [30] to vision transformers like TransReID [4] and PASS [31]. Even pre-trained on a large-scale unsupervised person dataset [30], LUP and PASS with OFTP still achieve mAP improvements of 2.6% and 1.0%, respectively, on the large-scale and challenging MSMT17 dataset. On CUHK03, which contains fewer training samples, BoT and LUP achieve even larger gains of 6.4% and 3.4%, highlighting the effectiveness of Reffusion and OFTP in low-data regimes. These results demonstrate strong generalization ability of our method across architectures and datasets, a property not shown in prior works.

As shown in Table III, we compare OFTP with state-of-the-art methods on person re-identification. OFTP improves ReID performance even when applied to the strong baseline BoT [3], whereas prior state-of-the-art methods have not demonstrated this capability, limiting their practical application. Moreover, training with Reffusion-generated samples substantially outperforms those from PIDM [12] and CFLD [13] under identical settings (except synthetic images), indicating that Reffusion produces more realistic person images which better support ReID training. In contrast, samples from PIDM contribute only marginal improvement, while samples from CFLD even degrade BoT.

D. Ablation Study and Analysis

The influence of components in Reffusion. We explore the effect of key modules in Reffusion in Table IV and Figure 4.

TABLE II

MAP IMPROVEMENTS OF OFTP ON DIFFERENT DATASETS AND BACKBONES. **BASELINE** INDICATES TRAINING SOLELY ON REAL DATASET, WHEREAS **W/SYNTHESIS.** REFERS TO TRAINING WITH ADDITIONAL GENERATED IMAGES.

Methods	Market1501		CUHK03		MSMT17	
	Baseline	w/ Synthesis.	Baseline	w/ Synthesis.	Baseline	w/ Synthesis.
BoT [3]	85.9	87.1+**1.2**	61.3	67.7+**6.4**	54.3	55.9+**1.6**
BoT+LUP [30]	87.5	88.5+**1.0**	69.5	72.9+**3.4**	54.6	57.2+**2.6**
TransReID [4] w/o SIE	88.2	89.3+**1.1**	78.6	80.8+**2.2**	67.1	68.3+**1.2**
PASS [31]	92.8	93.4+**0.6**	88.7	89.4+**0.7**	74.2	75.2+**1.0**

TABLE III

COMPARISON OF MAP WITH STATE-OF-THE-ART METHODS ON PERSON RE-IDENTIFICATION. MOST PRIOR WORKS ADOPT BOT [3] BACKBONE BUT WITH DIFFERENT STANDARDS. * INDICATES THE USE OF OFTP.

Datasets	Methods	Baseline	w/ Synthesis
Market1501 (256×128)	PIDM [12]	76.7	78.4
	PCDMs [14]	76.7	80.3
	Diff-ID [13]	68.5	72.6
	Identity Diffuser [16]	81.6	84.2
	PIDM* [12]	85.9	86.2
	CFLD* [13]	85.9	85.0
	Reffusion* (Ours)	**85.9**	**87.1**

TABLE IV

ABLATION STUDY OF THREE KEY COMPONENTS IN REFFUSION ON MARKET1501 WITH RESPECT TO PERSON IMAGE QUALITY.

Methods	Variants	FID↓
B1	`Baseline`	9.657
B2	`Baseline+AMT`	9.033
B3	`Baseline+AMT+OMM`	8.823
Ours	`Baseline+AMT+OMM+IGG`	**8.008**

TABLE V

ABLATION OF OMM REGULARIZATION ON MARKERT1501.

Methods	Variants	FID↓
C1	`Baseline+AMT+OMM w/o` $\mathcal{L}_{\text{reg}}$	9.104
C2	`Baseline+AMT+OMM`	**8.823**

In Table IV, FID gradually decreases as more modules are incorporated. The `Baseline` relies solely on global semantic features. AMT and OMM enable Reffusion to better focus on synthesis of human textures (as illustrated in row 1 and 2 of Figure 4, comparing B1 with B2 and B3). IGG further reduces FID score by 0.815, enabling generated samples to more closely resemble the source person image, particularly when Reffusion fails to accurately capture clothing details like color information (see the last column of row 3 of Figure 4).

The regularization of OMM. As presented in Table V, without uniformity regularization for memory diversification, additional parameters in OMM does not lead to improved FID performance, which demonstrates the effectiveness of $\mathcal{L}_{\text{reg}}$.

The effect of OFTP. Table VI shows the effectiveness of proposed OFTP. It can be seen from index 0 that directly combining generated samples with original training dataset yields suboptimal improvement compared to baselines in Table II. By employing k-nearest strategy, a generated image that doesn't share the same identity as any of its k-nearest centroid prototypes is filtered out, leading to mAP improvement (index 1 in Table VI). Increasing k helps mine hard positive samples. However, if k is too large, noisy and poor-quality samples may be included, hinder the post-training of ReID model (index 2 in Table VI). This issue can be mitigated by applying a threshold τ to corresponding similarity. In this way, the impact of low-quality cases during hard generated sample mining is reduced, resulting in additional +1.0% mAP improvement (index 4 in Table VI), compared to the setting without any filtering strategy. The hyper-parameter τ is adaptively determined by the mean similarity between real samples and their centroid prototypes in our experiments, where such similarity reflects intra-class compactness in the hypersphere. And we also find that using 10-nearest neighbors works well.

Fig. 4. Qualitative ablation results of components in Reffusion.

Efficiency and convergence. Reffusion has a comparable training time to LDM-based [32] methods. By removing VAE, Reffusion avoids the inference overhead introduced by latent encoding, and adopts a much lighter UNet (170M vs. 860M in LDM), without increasing computational cost significantly. Moreover, Reffusion converges faster and more stably: while LDM models often suffer from loss saturation—mainly due to VAE-induced detail degradation in low-resolution and blurry persons, Reffusion stably reduces loss with moderate training.

IV. CONCLUSION

We propose Reffusion, a diffusion-based framework for person image synthesis, together with an Online Filtering and

TABLE VI
ABLATION STUDY OF OFTP ON DIFFERENT DATASETS WITH BOT BACKBONE.

Index	thresholding	k-nearest	mAP			
			Market1501	CUHK03	MSMT17	Average
0	✘	✘	86.0	66.1	55.6	69.23
1	✘	1-nearest	86.8	67.4	55.7	69.97
2	✘	5-nearest	86.4	67.7	55.5	69.87
3	✓	1-nearest	86.7	67.3	55.4	69.80
4	✓	5-nearest	**87.1**	**67.7**	**55.9**	**70.23**
5	✓	10-nearest	87.2	67.4	55.9	70.17

Training Policy (OFTP) for ReID post-training. Firstly, Reffusion generates high-fidelity person images with strong identity preservation, accurate pose transfer and reduced background interference through OMM, AMT and IGG modules. Then, OFTP improves ReID performance by adaptively filtering out low-quality generated samples and preserving informative ones for post-training. Extensive experiments on Market1501, MSMT17 and CHUK03 datasets demonstrate the superiority and generalization of our proposed Reffusion and OFTP.

REFERENCES

[1] W. Zajdel, Z. Zivkovic, and B. J. Krose, "Keeping track of humans: Have i seen this person before?" in *Proceedings of the 2005 IEEE international conference on robotics and automation*. IEEE, 2005, pp. 2081–2086.

[2] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," 2018, pp. 274–282.

[3] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," pp. 2597–2609, 2019.

[4] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 013–15 022.

[5] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," 2018, pp. 480–496.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[8] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 650–667.

[9] P. Zhang, L. Yang, J.-H. Lai, and X. Xie, "Exploring dual-task correlation for pose guided person image generation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022, pp. 7713–7722.

[10] Y. Ren, X. Fan, G. Li, S. Liu, and T. H. Li, "Neural texture extraction and distribution for controllable person image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 535–13 544.

[11] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[12] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah, and F. S. Khan, "Person image synthesis via denoising diffusion model," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5968–5976.

[13] Y. Lu, M. Zhang, A. J. Ma, X. Xie, and J. Lai, "Coarse-to-fine latent diffusion for pose-guided person image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6420–6429.

[14] F. Shen, H. Ye, J. Zhang, C. Wang, X. Han, and W. Yang, "Advancing pose-guided image synthesis with progressive conditional diffusion models," *arXiv preprint arXiv:2310.06313*, 2023.

[15] F. Shen and J. Tang, "Imagpose: A unified conditional framework for pose-guided person generation," *Advances in neural information processing systems*, vol. 37, pp. 6246–6266, 2024.

[16] S. Nie, Z. Shi, R. Liu, S. Guo, M. Zhang, M. Wang, K. Osamura, L. Septiana, and A. Narishige, "Attribute conditional diffusion-augmented person re-identification," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[17] I. H. Kim, J. Lee, S. Son, W. Jin, K. Cho, J. Seo, M. Kwak, S. Cho, J. Baek, B. Lee *et al.*, "Pose-diversified augmentation with diffusion model for person re-identification," *CoRR*, 2024.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[19] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4D: Reconstructing and tracking humans with transformers," in *ICCV*, 2023.

[20] N. Sloane, R. Hardin, W. Smith *et al.*, "Spherical codes," *URL http://www2. research. att. com/˜ njas/packings*, 2000.

[21] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International conference on machine learning*. PMLR, 2020, pp. 9929–9939.

[22] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[23] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016, pp. 770–778.

[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," 2015, pp. 1116–1124.

[26] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," 2018, pp. 79–88.

[27] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," 2014, pp. 152–159.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[30] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang, H. Li, and D. Chen, "Unsupervised pre-training for person re-identification," 2021, pp. 14 750–14 759.

[31] K. Zhu, H. Guo, T. Yan, Y. Zhu, J. Wang, and M. Tang, "Pass: Part-aware self-supervised pre-training for person re-identification." Springer, 2022, pp. 198–214.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.