Privacy Preserving Person Re-identification via Anonymizing Diffusion Model

Anonymous Authors

ABSTRACT

Person re-identification (ReID) has made impressive progress in recent years, yet the associated privacy issues received scant attention. For example, an attacker may use social media images to retrieve relevant pedestrians from a database to obtain private information. To address this issue, numerous anonymization methods have been proposed to reconstruct datasets, such as image encryption and adversarial attack. However, existing anonymization methods do not change the person IDs, thus leaving privacy vulnerabilities as the true IDs can still be retrieved through decryption or human observation. In this paper, we propose an anonymizing diffusion model (ADM), a novel generation-based privacy protection approach for ReID. ADM utilizes stable diffusion to generate anonymous images through text prompts, preventing being retrieved by either neural networks or human observers. Furthermore, we introduce a new metric called ID separation degree (ISD) to quantitatively measure the visual difference between real and generated samples, reflecting the anonymization level. To the best of our knowledge, this is the first exploration of utilizing diffusion models to generate new IDs for privacy protection ReID. Experiments on public datasets demonstrate that ADM achieves excellent privacy protection performance while maintaining competitive ReID accuracy. The source code is available at https://anonymous.4open.science/r/ADM-2423

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**.

KEYWORDS

person re-id, privacy protection, stable diffusion

1 INTRODUCTION

In recent years, with the rapid development of artificial intelligence technologies, personal privacy protection has also received increasing attention. Some institutions and research projects have taken corresponding measures to mitigate privacy leakage risks. For example, to avoid public disclosure of identification information, famous datasets such as DukeMTMC [46] and Tiny ImageNet [20] have been modified or withdrawn in subsequent versions. The ImageNet dataset has also added blurring processing to images containing facial features. Technology companies such as Meta have also banned the use of facial recognition systems for privacy reasons. These measures highlight the inherent contradiction between utilizing personal images for AI research and protecting privacy.

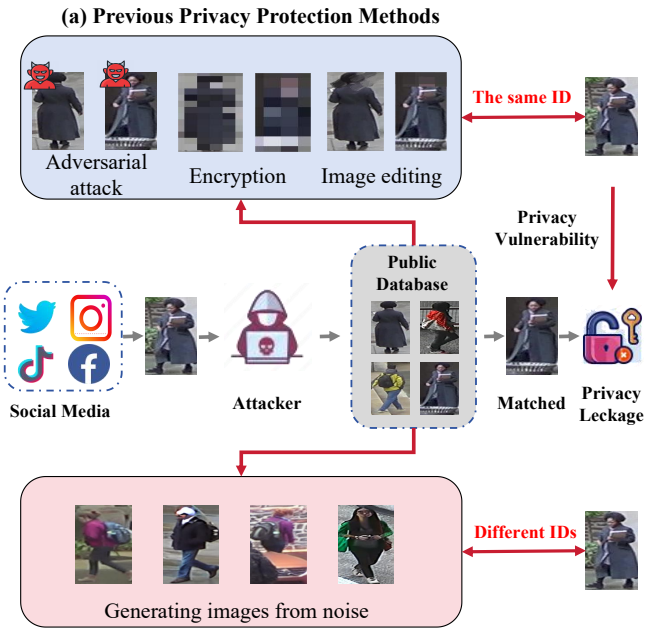

Figure 1: The difference between ADM and other methods. (a) The previous methods protect ID privacy through attacking, encryption, editing, etc. These methods fail to erase the correlation with the original ID, leaving private information that can still be retrieved through human observation or decryption. (b) ADM (ours) protects ID privacy by generating new IDs to replace the original ones.

As a significant subtask of retrieval, person re-identification (ReID) faces severe privacy leakage issues. The goal of ReID technology is to detect and track the same target person across different camera views based on image features. Mainstream ReID datasets such as Market-1501 [45], DukeMTMC-reID, CUHK03 [16] and MSMT17 [35] are mostly collected from public scenarios such as shopping malls and campuses, but whether the collection of these datasets has fully considered privacy protection of the participants remains in doubt. As shown in Figure 1 (a), as long as obtaining a photo of these participants from social media, attackers may attempt to retrieve images from public datasets for more privacy information to commit crimes. Taking DukeMTMC-reID as an example, it contains a large amount of unauthorized image resources, which triggered the problem of user privacy leakage and led to the withdrawal of this dataset later. Although these datasets have made

Unpublished working draft. Not for distribution.

ACM MM, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

important contributions to the advancement of ReID research, insufficient privacy protection also hinders the landing of the technology in real application environments.

To address this issue, some existing works on privacy protection for ReID attempt to reconstruct datasets using image encryption or adversarial attacks. Image encryption aims to selectively blur sensitive regions [7], adding noise to the image [19], or encrypt [42] before releasing datasets. The images obtained by encryption-based methods require a complex decryption process before being utilized by ReID models, which limits the application scope of encryption-based methods. Adversarial attack approaches [18, 32, 33] add carefully crafted perturbations to samples to reduce their recognizability during retrieval.

However, to ensure data usability, these methods do not change the IDs of pedestrians. For example, adversarial attack methods only add slight perturbations to the images, allowing the true IDs to still be retrieved by humans or other ReID models. Similarly, image encryption methods can be decrypted to obtain the original image. Essentially, these anonymization methods hide private information rather than eliminate it. Therefore, these methods still have the risk of privacy leakage.

To eliminate the privacy information of the original data, we propose an anonymizing diffusion model (ADM), a novel generation-based privacy protection approach for ReID. ADM aims to anonymize ReID datasets by generating visually distinct new IDs, preventing being retrieved by either neural networks or human observers. To this end, ADM utilizes Stable Diffusion (SD) [26] to generate anonymous images with new IDs through text prompts. As shown in Figure 1, compared with previous methods, ADM is a more thorough solution for privacy protection with wider applications. Furthermore, to enhance the anonymization and utility capability, ADM leverages identity text descriptions as prompts to control SD generation, ensuring inter-identity variability and intra-identity consistency. Specifically, ADM can be divided into three steps: 1) Perform 2-stage fine-tuning of the SD model using these pairs. 2) Construct novel text prompts and generate a new anonymous dataset. 3) Filter outlier samples from the generated dataset.

Additionally, to quantitatively measure the visual difference between original and generated samples, we introduce a new metric called ID separation degree (ISD). ISD measure anonymization capability. of a privacy protection method based on the feature distance in the embedding space. A higher ISD indicates greater dissimilarity between the real and generated samples, and thus enhanced privacy protection level.

To analyze the effectiveness of our proposed method, we conduct thorough experiments on commonly used ReID datasets and compare them with other privacy-preserving methods. The results show ADM achieves both competitive utility capability and privacy-preserving capability.

In summary, the contributions of our work can be concluded as:

1) We propose a novel generation-based privacy protection approach for ReID called ADM, preventing original IDs from being retrieved by either neural networks or human observers.

2) We introduce a new metric called ID separation degree (ISD) to quantitatively evaluate the privacy protection capability, through measuring the visual difference between real and generated samples.

3) We conduct experiments on benchmark datasets to demonstrate that ADM can effectively remove identity associations between real and generated data, with superior performance in preventing retrieving private identities from public datasets, while maintaining competitive ReID accuracy.

2 RELATED WORK

2.1 Person Re-Identification

Person re-identification (ReID) is an important person retrieval task, which aims to retrieve a person of interest across multiple non-overlapping camera views. It has wide applications in video surveillance, intelligent security, and other fields. In recent years, many deep learning-based methods have been proposed to solve this problem. CNNs as the backbone network [36, 39, 40, 44] are widely used for feature extraction of pedestrian images. Metrics-based learning methods are commonly used to calculate the similarity between person images [4, 22, 31]. Deng et al. [5] apply GAN to ReID to solve the domain adaptation problems. Some powerful baselines have also been proposed, such as AGW [41], which has had a wide and profound influence on the ReID community. However, person ReID still faces severe privacy issues. For example, the widely used dataset DukeMTMC-reID [46] was withdrawn due to privacy issues. Most existing ReID models rarely consider privacy protection problems, and some methods proposed have difficulty in balancing privacy protection and model performance

2.2 Privacy-Preserving Methods

Traditional privacy-preserving methods focus on blurring, pixelating, or adding noise to images, which may affect the ReID performance to some extent. Dietlmeier et al. [6] show that blurring faces has little impact on the performance of ReID systems, and propose an anonymous ReID dataset DAA [7] with blurred faces. However, the face only occupies a small part of the pedestrian image, and privacy attackers can still retrieve the target person through other major features such as clothing, posture, background, etc. Zhang et al. [43] propose a reversible anonymous framework based on joint learning. They adapt desensitized images generated by conventional methods as the initial supervision to generate anonymous images for privacy protection, and the original images can be recovered for ReID research. PIS [8] uses images from other identities in the dataset to weakly encrypt the original image, generating anonymized images with relatively consistent identities. Differential privacy [9, 10] and secure computation [13, 38] have also been used for privacy protection, yet they introduce relatively high computational overhead [15]. Some synthetic virtual datasets [3, 28, 34] have better privacy protection effects while the large domain gap results in poor generalizability. Recently, event cameras have also been used for privacy-preserving in ReID [1, 2], but their application is not widespread.

2.3 CLIP and Stable Diffusion

CLIP [24], as a pre-trained model based on a transformer architecture, has been widely applied to various multimodal tasks [12, 23, 29, 30]. Through contrastive learning on a huge dataset, CLIP builds semantic connections between text and images to achieve mutual

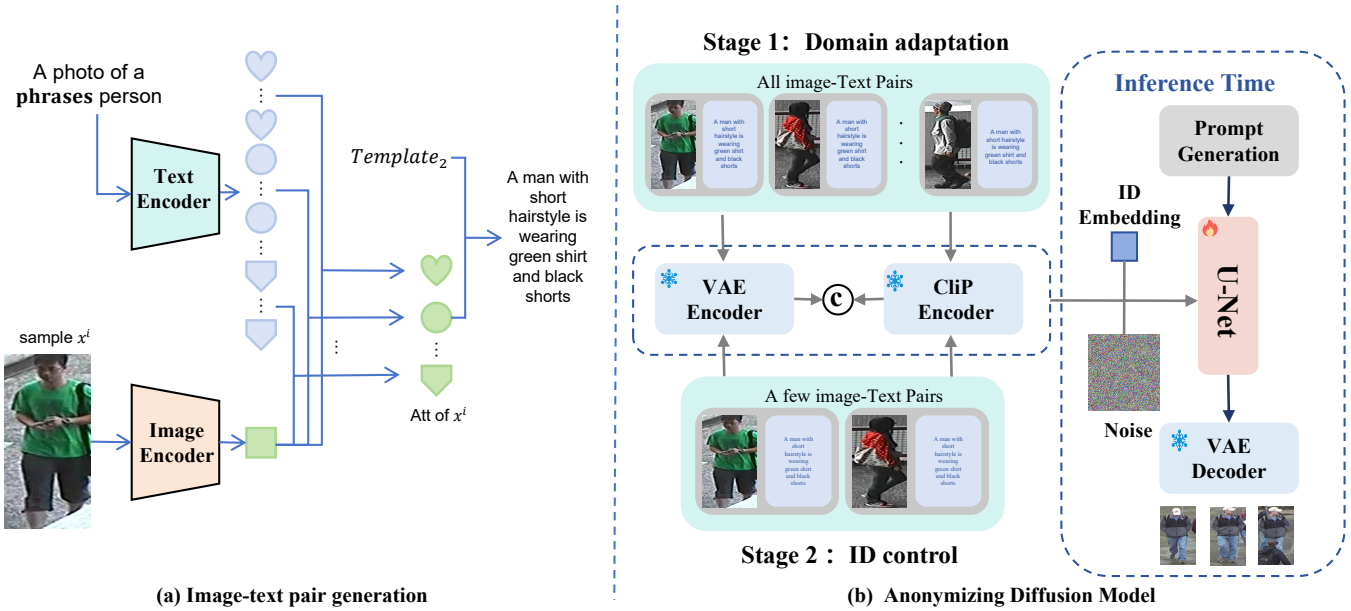

Figure 2: Illustration of Image-text pair generation and ADM. (a) We use CLIP to encode both prompts and images, then use cosine similarity to match the most fit attributes to the image. The attributes selected are then fitted into a template as a text caption of the image. (b) We fine-tune SD in 2 stages. We use the whole image-text pairs in Stage 1 and a small subset in Stage 2 to fine-tune the U-Net, with other parts frozen. We add ID (class) Embedding to the time step embedding to improve the intra-identity consistency.

understanding between images and text. ClipCap [21] uses CLIP encodings as image caption prefixes, and fine-tunes a language model GPT2 [25] to generate image captions through a simple mapping network. Shao et al. [27] utilize a divide-conquer-combine strategy to generate text descriptions of images through the CLIP paradigm, constructing a large-scale text-labeled person dataset "LUPerson-T". encoder of CLIP to generate prompts' latent embeddings as conditions to guide image generation. Diffusion models [14] are generative models that include a process of forward noise injection diffusion and a reverse process of denoising reconstruction. An improvement of Stable Diffusion [26] over Diffusion models is that it performs diffusion in the latent space of images, achieving higher computational efficiency. In addition, Stable Diffusion introduces conditional control. It uses the text encoder of CLIP to generate prompts' latent embeddings as conditions to guide image generation.

3 METHODOLOGY

The goal of ADM is to generate new IDs that are anonymized and ReID usable. In the following, we describe the main components of our method. 1) Image-text Pair generation (Sec. 3.1): As shown in Figure 2 (a), We use CLIP [24] and cosine similarity to generate text prompts from real data samples as image-text pairs. 2) Anonymizing Diffusion Mode (Sec. 3.2): As shown in Figure 2 (b), the SD is fine-tuned using image-text pairs to learn the Re-ID image style and ID consistency. Then, we use fine-tuned SD to generate new IDs with new descriptive prompts. To refine the generated dataset, we use ADM Griddle to filter the inconsistent samples for a given ID. 3) ID separation degree (Sec. 3.3): To evaluate the anonymization capability of ADM, we introduce a new metric called ID separation degree (ISD).

3.1 Image-text Pair Generation

Previous work solely relying on noise to generate new images cannot ensure consistency between identities, which is crucial for ReID training. Therefore, we use text to control identity consistency. We adopt an automatic img2text generation approach since manually annotating image-text pairs is time and labor-costly. However, directly using CLIP to generate captions for ReID images has limitations as it struggles to produce fine-grained, consistent attribute descriptions necessary to effectively control the image generation process for ReID tasks. Inspired by T2I-ReID [27], we implemented img2text generation using the Divide-Conquer-Combine approach, as shown in Figure 2 (a).

Divide. Based on the experience from T2I-ReID, we initialize a set of attribute description terms:

$$\begin{aligned} Att \in \{ &< gender >, < hair >, \\ &< upper\_wear\_color >, < upper\_wear >, \\ &< lower\_wear\_color >, < lower\_wear > \} \end{aligned} \tag{1}$$

Each attribute Att is a set of description terms corresponding to it, e.g. $< gender >$ contains the terms *man* and *woman*. These attributes cover the basic appearance of pedestrians and can provide sufficient information to train a ReID model.

Conquer. Denote the set of samples in the original dataset as $X = \{x^i | i \in (0, n)\}$. We populate template $Template_1$ with the attributes mentioned earlier to obtain the following set of sentences:

$$T_{Att} = \{t^j_{Att} | t^j_{Att} = \text{"A photo of a } \textbf{phrase} \text{ person"}, \text{ where } \textbf{phrase} \in Att\} \quad (2)$$

Encoding x^i and t^j_{Att} through CLIP gives f^i_x and f^j_{Att}. We calculate the cosine distance between f^i_x and each f^j_{Att}, and select the attribute term with the maximum value, which can be denoted as:

$$p^i_{Att} = \arg\max_j \cos\left(f^i_x, f^j_{Att}\right) \quad (3)$$

Combine. Each sample x^i after the Conquer phase will obtain a set of description terms: p^i_{Att}. We populate these description terms into $Template_2$ to obtain the caption for x^i:

$$\begin{aligned} Caption = &\text{A} < gender > with < hair > \text{is wearing} \\ &< upper\_wear\_color > < upper\_wear > \text{and} \\ &< lower\_wear\_color > < lower\_wear > . \end{aligned} \quad (4)$$

The above img2text model can generate captions suitable for ReID. Note that different from T2I-ReID, we only use a fixed Template 2 for new prompt generation, and do not select optional attributes like $< bag >$, $< hat >$, $< vehicle >$ to focus on identities.

3.2 Anonymizing Diffusion Model

3.2.1 ***Stable Diffusion Fine-tuning.*** ReID model is sensitive to domain shift and intra-identity consistency. Domain gap between generated training data and real test data should be eliminated to achieve satisfying performance. Samples belonging to the same ID should have the same pedestrian visual features so that the model can learn to extract robust feature representation for re-identification. However, existing SD models hardly match these characteristics due to their pre-training data. To generate images close to original dataset in style and consistency, we apply 2-stage fine-tuning to SD, as shown in Figure 2 (b) To further enhance the consistency of discriminative features for same-ID pedestrian images, we introduce class embeddings.

Stage 1. The main goal of stage 1 is to fine-tune the style of generated images to match those in original datasets and develop the preliminary ability to generate realistic and discriminative pedestrian images conditioned on prompts. To achieve this goal, we fine-tune the pre-trained model Stable-Diffusion-v-1.5 on the entire original training dataset containing real image-text pairs from Sec. 3.1. SD typically produces high-quality images with clear resolution without ReID-specific characteristics like low resolution, variational illumination and backgrounds. Fine-tuning on real images allowed the model to learn these unique ReID styles and nuances. Additionally, by training on pairs of images and corresponding text, the model preliminarily learned to match prompt descriptions and visual attributes and features of pedestrians in the images.

Through this first-stage fine-tuning process, we assimilated the style and captured basic text-conditioning abilities without overfitting to specific identities, laying the foundation for subsequent controlled generation tailored to the ReID task.

Stage 2. In the second stage of fine-tuning, the main goal was to map prompts more precisely to pedestrian features to better control the SD generation process. After stage one, SD could not accurately interpret prompt terms due to overlapping attributes across identities in the full dataset. Also, prompt can hardly describe the detailed appearance and texture of an image precisely. To tackle this problem, attributes in prompt should be bound to certain visual patterns in the images.

Stage two addressed this by fine-tuning SD on a small selected subset from the first stage. This refined generation so that within-ID consistency improved and features conformed more closely to prompts. Specifically, we cropped a small portion of image-text pairs while ensuring the proportions of a certain attribute did not exceed a threshold. This prevents duplicate attributes describing different identities, allowing SD to learn direct mappings from prompt to pedestrian appearance. The refined generation produced by stage 2 therefore featured higher conformity between prompts and generated identity features for more realistic augmented data.

Class Embedding. When training SD models, we initialize a unique learnable class embedding emb^i_{cls} for ID id^i in training set. We use $emb^i_{cls} + emb_{time}$ to replace the original time step embedding emb_{time} of SD. Compared to using emb_{time} alone, our method can provide more continuously consistent contextual information, thereby mitigating consistency issues caused by random noise. This additional conditioning helps produce even more consistent results.

3.2.2 ***New Prompt Generation.*** In this section, we will generate prompts for new IDs. It is intuitive to use existing prompts, but this may cause a lack of diversity and privacy protection ability. This is because directly using the original text prompts may easily associate the generated images to existing IDs since we fine-tuned SD using image-text pairs. Therefore, we need to use text prompts not appearing in the subset we selected in Sec. 3.2.1 to generate new IDs. We propose a method to automatically construct richer and more discriminative prompts. We categorize the attributes generated in Sec. 3.1 into three groups:

$$\begin{cases} p_h = < gender > + < hair > \\ p_u = < upper\_wear\_color > + < upper\_wear > \\ p_l = < lower\_wear\_color > + < lower\_wear > \end{cases} \quad (5)$$

Suppose the subset selected in Sec. 3.2.1 contains m different IDs $x^1, x^2, \cdots, x^m$, where the attribute group of each $x^i (i = 1, 2, \cdots, m)$ is $p^i = (p^i_h, p^i_u, p^i_l)$. Theoretically, we can combine these m attribute groups into m^3 prompts. Let the set of all m^3 prompts be P_{all} and m prompts of selected IDs be P_{orgn}.

We use Algorithm 1 to select prompts P_{remain} from P_{all}. The algorithm ensures all prompts in P_{remain} have almost different attributes, preventing different generated IDs from sharing similar prompts. Thus the diversity of generated dataset is further ensured.

We input the attribute groups generated by the above algorithm into equation 4 as prompts for the fine-tuned SD model. This allows the model to generate a novel dataset with the potential for new IDs. By feeding these prompts into the fine-tuned SD model, we can gain additional synthetic pedestrian images never seen during training. The new dataset thus achieves the goals of reconstructing identity features for ReID while avoiding memorizing sensitive

Algorithm 1 Prompt Griddle

function GriddleIter($P_{all} : list, P_{orgn} : list, n : int$)
  $check_u, check_l, P_{select}, backup \leftarrow$ empty list
  for $i = 0$ to $len(P_{all})$ **do**
    if $p^i \notin P_{orgn}$ **then**
      if $p_u^i \notin$check$_u$ & p$_l^i \notin$check$_l$ **then**
        Add p^i to P_{select}
        Add p_u^i to $check_u$
        Add p_l^i to $check_l$
        if $len(P_{select}) >= n$ **then**
          return $P_{select}, backup$
      else
        Add p^i to $backup$
  return $P_{select}, backup$

procedure PromptGriddle($P_{all} : list, P_{orgn} : list, n : int$)
  $P \leftarrow P_{all}, P_{remain} \leftarrow$ empty list, $num \leftarrow n$
  while $len(P_{remain}) < n$ **do**
    $P_{select}, backup = GriddleIter(P, P_{orgn}, num)$
    Add P_{select} to P_{remain}
    $P = backup$
    $num = num - len(P_{select})$
  return P_{remain}

private training data details, fulfilling both data utility and privacy preservation.

3.2.3 ***ADM Griddle.*** Through the former process, we have obtained a novel dataset with reasonable intra-identity consistency by controlling generation with text prompts. However, inconsistency can still occur due to noise introduced from the original dataset and fine-tuning process. To ensure the potential ReID performance of the generated dataset is not adversely affected, we need further screening to guarantee consistency within each ID. We propose a new method called ADM Griddle to filter inconsistent samples for a given ID. Let $\{x^1, x^2, \cdots, x^n\}$ and $\{F^1, F^2, \cdots, F^n\}$ denote samples and their features of a specific id. We specify a hyperparameter ϵ, and the set of samples within a distance of ϵ from x^i is $S(x^i)$. We choose the ID with the largest set,

$$k = \arg\max_{i \in (0,n)} |S(x^i)| \tag{6}$$

ADM Griddle retains $S(x^k)$ and filters out the rest. ADM Griddle removes outlier samples that do not closely match the densest cluster of samples for each identity, thereby enhancing the intra-identity consistency of the filtered dataset.

3.3 ID separation degree (ISD)

Anonymized images should not be visually similar to any real image for privacy, nor should they have similar discriminative features to real images. Generated identities should be mutually independent from real ones. Thus, we can measure their similarity by the distance between feature representations of generated identity images and real identities. Anonymized features should be as far as possible from real identity features in the feature space to present distinct visual characteristics and achieve anonymization. To characterize the visual difference between samples, we propose a new metric called ID separation degree (ISD). Given a real dataset with n identities denoted as $id^1, id^2, \cdots, id^n$, and the number of samples belonging to each id^i is m^i. The k-th sample of the i-th identity in the feature space is f_k^i The centroid of the i-th identity is:

$$c^i = \frac{1}{m^i} \sum_k f_k^i \tag{7}$$

The generated data contains $m_{\mathcal{G}}$ samples with features $f_{\mathcal{G}}^j$, where $j \in [1, m]$. ISD is defined as:

$$\text{ISD} = \frac{1}{m_{\mathcal{G}}} \sum_j \min_i \left[1 - \frac{\|c^i \times f_{\mathcal{G}}^j\|}{\|c^i\| \cdot \|f_{\mathcal{G}}^j\|} \right] \tag{8}$$

ISD uses the cosine distances between generated sample features and the centroid of real identities. A higher ISD indicates lower similarity between real and generated data in the feature space on average, hence stronger privacy protection by preventing identity association. ISD effectively evaluates the degree of anonymity in privacy-preserving ReID methods.

4 EXPERIMENT

To analyze the effectiveness of our proposed method, we conduct thorough experiments on commonly used ReID datasets. The remainder of this section presents the implementation details of our framework (Sec. 4.1), datasets we conduct experiments on (Sec. 4.2), results of experiments (Sec. 4.3), ablation studies (Sec. 4.4).

4.1 Datasets

We conduct experiments on the two commonly used ReID datasets: Market-1501 [45], DukeMTMC-reID [46], and CUHK-SYSU[37]. The details of the two ReID datasets are as follows:

1) Market-1501 contains 12,936 training images of 750 IDs, 3368 query images of 750 IDs, and 15,913 gallery images of 751 IDs. All the images are from 6 cameras on the campus of Tsinghua University.

2) DukeMTMC-reID contains 16,522 training images of 702 IDs, and 2228 query images of 702 IDs. All the images are from 8 cameras on the campus of Duke University.

3) CUHK-SYSU contains 11,206 training images of 5,532 IDs, and 6,978 query images of 2,900 IDs. Within these images, 12,490 of 6,057 IDs are collected with hand-held cameras across streets, and 5,694 of 2,375 IDs are from movies and TV dramas.

4.2 Implementation Details

SD Fine-tuning. We use Stable-Diffusion-v-1.5 for fine-tuning. The batch size is set to 48. To optimize the model, we use Adam optimizer with a learning rate of 10^{-4} and weight decay of 10^{-2}. The training state takes 100 epochs in Stage 1, and 50 epochs in Stage 2. Within the subset, there are 3 IDs for each group, 25 images for each ID, and we set the portion of overlap Att to 0.4.

ReID Training. We use Resnet50 for the backbone. The batch size is set to 64. To optimize the model, we use Adam optimizer with base learning rate 3.5×10^{-4} and weight decay of 5×10^{-4}. The total training state takes 120 epochs.

Table 1: Compare with other methods. We divide ADM and existing methods into two settings and conduct experiments on Market-1501 and DukeMTMC-reID. Under the first setting, methods only aim to improve its ReID performance through generative UDA. Under the second setting, methods aim to achieve a trade-off between ReID performance and privacy protection. The R-1, R-5, R-10 indicats the metric Rank-1, Rank-5, Rank-10 respectively. 1x/2x data denote that the amount of generated IDs and images are the same as/twice that of the real training set.

Methods	Settings	Market-1501 [45]				DukeMTMC-reID [46]			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
PTGAN [11]	Generative UDA	38.6	-	66.1	-	27.4	-	50.7	-
SPGAN [5]		51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
CamStyle [47]		58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
PersonX Subl [28]	Privacy Protection	34.8	46.1	51.6	17.2	22.4	29.8	34.6	13.4
PersonX Sub4 [28]		48.6	61.4	67.2	29.9	33.1	42.2	47.7	21.7
TRS [17]		64.6	74.9	79.5	44.7	50.1	61.3	65.9	**39.3**
ADM(1 x data)		67.4	83.8	88.4	45.3	52.3	68.9	74.0	36.0
ADM(2 x data)		**72.1**	**87.2**	**91.4**	**50.6**	**54.7**	**70.4**	**75.76**	36.3

Table 2: Results on different datasets. High T_g and T_u indicate great utility capability. High ISD, and low T_p indicate great privacy-preserving capability.

Datasets	T_g				T_u				T_p				ISD		
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	Orgn	ADM	Impv
Market-1501 [45]	67.4	83.8	88.4	45.3	94.8	97.9	98.7	72.5	57.1	68.4	73.6	19.4	0.15	0.50	2.33
DukeMTMC-reID [46]	52.3	68.9	74.0	36.0	95.2	98.6	99.0	78.8	71.3	82.3	85.4	19.9	0.11	0.53	3.82
CUHK-SYSU [37]	72.0	83.6	86.5	68.7	94.8	98.1	98.7	96.2	21.7	31.9	37.1	6.6	0.13	0.63	3.85

Figure 3: A demo of generated images. We generate images on three datasets. In each row, the two images on the right are generated from the three images on the left.

4.3 Experimental Results

4.3.1 ***Comparison With State-of-the-Arts.*** As shown in Table 1, we compare ADM with state-of-the-art methods under two settings, generative style transfer and privacy protection, on the Market-1501 and DukeMTMC-reID datasets. Methods under the first setting only aim to improve ReID accuracy through generative unsupervised domain adaptation (UDA). Methods under the second setting aim to achieve a trade-off between ReID performance and privacy protection.

Compared with the first setting methods, ADM achieves competitive results, indicating ADM is practical under real-world scenarios. When transferring to the target domain, generative style transfer methods leverage images without privacy protection, which may cause privacy issues. With ADM, ReID models can be transferred to the target domain using anonymized images without privacy concerns , while maintaining competitive performance.

Under the second setting, we compare ADM with privacy protection methods. On Market-1501, we obtain the best performance with **Rank-1=67.4, mAP=45.3**. Compared to PersonX Sub4[28],

we achieve 18.8 points and 15.4 points improvement on Rank-1 and mAP, respectively. Compared with the SOTA method TRS [17] based on virtual samples synthesis and data augmentation, we also achieve competitive results with 4.3% and 1.3% improvement on Rank-1 and mAP. On DukeMTMC-reID dataset, ADM achieved **52.3 on Rank-1 and 36.0 on mAP**, surpassing PersonX by a large margin. Compared with TRS, ADM gained a 4.4% improvement on Rank-1 and an 8.1% improvement on Rank-10. The performance of ADM is on par with that of TRS while adapting a much simpler framework without complex pose blending and 3D rendering. The experiment result indicates that ADM can generate images that much better fit the real-world data distribution than 3D models adapted in PersonX.

It is worth noticing that ADM can generate more IDs and images than real training sets. Recomposing m original prompts can bring at most $m^3 - m$ new prompts, thus more new IDs can be generated during new prompt generation. As shown in Table 1, on Market-1501, doubling the training data results in a 6.9% improvement on Rank-1 and an 11.6% improvement on mAP. By providing much more training data for the ReID model than normal datasets, ADM can further improve its ReID performance.

4.3.2 ***Results on benchmark datasets.*** To further quantitatively evaluate ADM, we conducted experiments on three benchmark datasets and evaluated utility capability and privacy capabilities using more metrics.

To evaluate the utility of the generated dataset, inspired by [8], we use two metrics, T_g and T_u. We trained a ReID model using the training set generated by ADM. Then we tested its mAP and rank-1 accuracy as T_p on the real test set and as T_u on the generated test set (not overlapping with the training set), respectively. For a fair comparison, the scale of the generated training set is consistent with the training set in the real dataset. T_g reflects the utility of generated data on training usable reID models targeting real scenarios. T_u measures discriminability within the generated data, considering inter-identity variability and intra-identity consistency.

From the perspective of privacy attacks, the attacker uses a real pedestrian image to retrieve related images in the target dataset. If the attacker successfully retrieves images related to that pedestrian, a privacy leakage has occurred. Therefore, we use the metric T_p corresponding to the mAP and Rank-1 of the retrieval results to measure the privacy-preserving capability during an attack through neural networks. A lower T_p indicates stronger resistance to attacks.

Furthermore, as mentioned in Sec. 3.3, we use ISD to measure the visual discrimination between generated data and original ones. A higher ISD indicates better anonymization capability of the model.

The results are shown in Table 2. Original (Orgn) of ISD denotes the average cosine distance between features of each original training sample and its ID's centroid. ADM of ISD denotes the average cosine distance between features of each generated training sample and its closest ID's centroid in the real training set. The improvement (Impv) of ISD is calculated as Impv = (ADM − Orgn)/Orgn, indicating the relative enhancement of privacy by ADM.

On each dataset, ADM achieves both high T_g and T_u, indicating that ADM is practically utilizable in ReID research. And simultaneously, ADM achieves high ISD indicating that the anonymized data are far dissimilar to the original data. The result of T_p is significantly lower than normal levels, which means that the ReID model fails to retrieve relevant IDs from the gallery protected by ADM.

Note that, compared with other datasets, ADM gets better results on CUHK-SYSU, which has far more IDs than other datasets. This indicates that ADM may benefit from a large amount of training IDs in real datasets since more real IDs lead to better diversity of generated data.

4.4 Ablation Studies

Table 3: Ablation study. To validate the effectiveness of the main components in ADM, we evaluate it on Market-1501 using (a) different fine-tuning strategies; (b) no class embedding; (c) different data filtering methods, and (d) the whole model ADM.

	Methods	T_g mAP	T_g R-1	ISD
(a)	only Stagel	16.97	38.42	**0.57**
	only Stage2	31.08	53.18	0.56
(b)	w/o class embedding	38.76	61.37	0.52
(c)	random	37.26	61.97	0.50
	centroid	45.14	**67.99**	0.49
(d)	ADM (full model)	**45.29**	67.37	0.50

4.4.1 ***Effectiveness of different components.*** We compare four experiments: (a) Different fine-tuning strategies (b) With/Without class embedding (c) Using different data filtering methods (d) Complete ADM:2-Stage fine-tuning + class embedding + ADM griddle.

The results are shown in Figure 3.

(a) vs (d): Single-stage fine-tuning achieves great privacy protection but poor utility. The first stage overfits duplicated attributes without precise semantics, which means prompts can not control the image generation process steadily and the appearance of generated images may be biased from prompts, leading to poor intra-identity consistency. The second stage lacks style from real data and has a nonnegligible domain gap with real data, causing a lack of reality.

(b) vs (d): Adaptation of class embedding significantly improves utility with minor privacy loss, validating its role in consistency. Class embedding can provide better control of generated images with the same ID, thus enhancing intra-identity consistency, which is fundamental in real-world applications.

(c) vs (d): Random griddle leads to notably lower utility than centroid/ADM griddle, which performs similarly. The result indicates that a carefully designed griddle can filter out outlier samples of each ID and further ensure intra-identity consistency. In the main experiment, we chose ADM griddle as it gives better results.

4.4.2 ***Demo of generated data.*** We use ADM to generate new IDs on three benchmark datasets. To visually present the anonymization capability of ADM, we present some demos of the generated IDs in Figure 3. For every dataset, the two images on the right in each row are generated from the three identities on the left. The generated IDs share similar styles with the original dataset. Furthermore, they have a very low visual association with the related

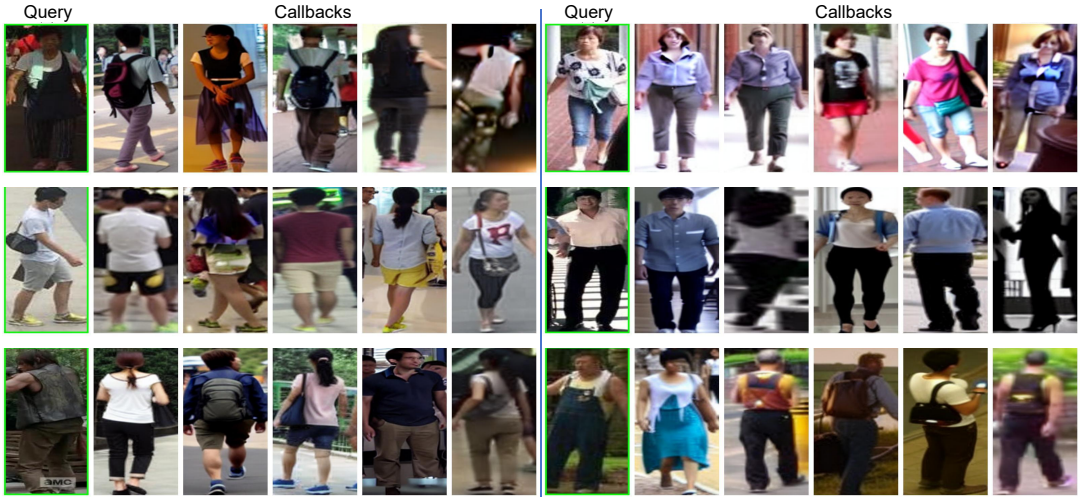

Figure 4: Visualization of attack results on the generated dataset. We use real CUHK-SYSU IDs to retrieve from the generated gallery protected by ADM. The callbacks are visually distinct from the real query, preventing privacy information leakage.

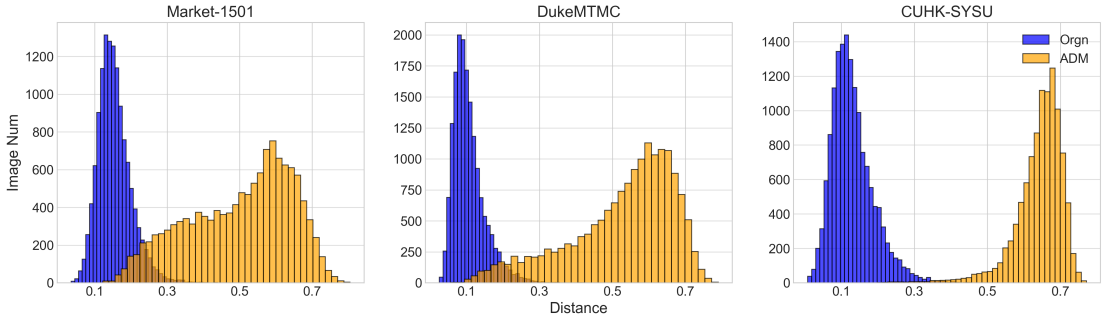

Figure 5: Cosine distance distribution.

original IDs, indicating that attackers would not be able to retrieve private information of the original IDs from the generated ones.

4.4.3 ***Demo of attack defense.*** To visually present the privacy-preserving capability of ADM, we visualize the retrieval result during attacks. We use real query images to retrieve generated images protected by ADM. As shown in Figure 4, images retrieved by the ReID model are not relevant to query IDs, thus the privacy attack is effectively resisted.

4.4.4 ***Cosine distance distribution.*** To gain deeper insights into the privacy protection differences among datasets, we analyze the distribution of 1) the distance between training images and the feature prototype(centroid) of their belonging IDs, and 2) the distance between generated images and the nearest real ID feature centroids. As shown in Figure 5, the distances between generated images and real ID centroids are significantly larger. This means that features of generated new IDs are distinctly different from real IDs, thereby achieving better privacy protection. Notably, the mean of ADM distribution is ISD. A more right-shifted ADM distribution indicates greater ISD.

5 CONCLUSION

In this work, we address the critical challenge of privacy protection for ReID datasets. We propose a novel approach called Anonymizing Diffusion Model (ADM) that uses stable diffusion to generate anonymous new identity images through text prompts. ADM prevents data from being retrieved by either neural networks or human observers while maintaining competitive ReID accuracy. Compared to prior approaches that rely on encryption or perturbations, ADM establishes a new standard for fully removing identity associations through controllable image synthesis. To quantitatively measure the visual difference between real and generated samples, we introduce a new metric ID Separation Degree (ISD). ISD allows comprehensive evaluation of anonymization through cosine distance between samples and centroids. Empirical results validate the effectiveness of ADM on benchmark datasets. We hope ADM can open up new research directions for privacy-preserving person re-identification.

REFERENCES

[1] Shafiq Ahmad, Pietro Morerio, and Alessio Del Bue. 2023. Person re-identification without identification via event anonymization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11132–11141.

[2] Shafiq Ahmad, Gianluca Scarpellini, Pietro Morerio, and Alessio Del Bue. 2022. Event-driven re-id: A new benchmark and method towards privacy-preserving person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 459–468.

[3] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. 2018. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*. 189–205.

[4] Kezhou Chen, Yang Chen, Chuchu Han, Nong Sang, Changxin Gao, and Ruolin Wang. 2018. Improving person re-identification by adaptive hard sample mining. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 1638–1642.

[5] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 994–1003.

[6] Julia Dietlmeier, Joseph Antony, Kevin McGuinness, and Noel E O'Connor. 2021. How important are faces for person re-identification?. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 6912–6919.

[7] Julia Dietlmeier, Feiyan Hu, Frances Ryan, Noel E O'Connor, and Kevin McGuinness. 2022. Improving person re-identification with temporal constraints. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 540–549.

[8] Shuguang Dou, Xinyang Jiang, Qingsong Zhao, Dongsheng Li, and Cairong Zhao. 2022. Towards privacy-preserving person re-identification via person identify shift. *arXiv preprint arXiv:2207.07311* (2022).

[9] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[10] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.

[11] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 4 (2018), 1–18.

[12] Kevin Frans, Lisa Soros, and Olaf Witkowski. 2022. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems* 35 (2022), 5207–5218.

[13] Craig Gentry. 2009. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 169–178.

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

[15] Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. 2020. Instahide: Instance-hiding schemes for private distributed learning. In *International conference on machine learning*. PMLR, 4507–4518.

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

[17] Yutian Lin, Xiaoyang Guo, Zheng Wang, and Bo Du. 2023. Privacy-protected Person Re-identification via Virtual Samples. *IEEE Transactions on Information Forensics and Security* (2023).

[18] Shaohao Lu, Yuqiao Xian, Ke Yan, Yi Hu, Xing Sun, Xiaowei Guo, Feiyue Huang, and Wei-Shi Zheng. 2021. Discriminator-free generative adversarial attack. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1544–1552.

[19] Kato Mivule. 2013. Utilizing noise addition for data privacy, an overview. *arXiv preprint arXiv:1309.3958* (2013).

[20] Mohammed Ali mnmoustafa. 2017. Tiny ImageNet. https://kaggle.com/competitions/tiny-imagenet

[21] Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734* (2021).

[22] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4004–4012.

[23] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2085–2094.

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

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

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

[27] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. 2023. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11174–11184.

[28] Xiaoxiao Sun and Liang Zheng. 2019. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 608–617.

[29] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. 2022. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–11.

[30] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2022. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3835–3844.

[31] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. 2018. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*. 365–381.

[32] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. 2020. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 342–351.

[33] Lin Wang, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. 2022. Attack is the best defense: Towards preemptive-protection person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*. 550–559.

[34] Yanan Wang, Shengcai Liao, and Ling Shao. 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.

[35] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 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.

[36] Dongming Wu, Mang Ye, Gaojie Lin, Xin Gao, and Jianbing Shen. 2021. Person re-identification by context-aware part attention and multi-head collaborative learning. *IEEE transactions on information forensics and security* 17 (2021), 115–126.

[37] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2016. End-to-End Deep Learning for Person Search. *CoRR* abs/1604.01850 (2016).

[38] Andrew C Yao. 1982. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*. IEEE, 160–164.

[39] Mang Ye, Cuiqun Chen, Jianbing Shen, and Ling Shao. 2021. Dynamic tri-level relation mining with attentive graph for visible infrared re-identification. *IEEE Transactions on Information Forensics and Security* 17 (2021), 386–398.

[40] Mang Ye, He Li, Bo Du, Jianbing Shen, Ling Shao, and Steven CH Hoi. 2021. Collaborative refining for person re-identification with label noise. *IEEE Transactions on Image Processing* 31 (2021), 379–391.

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

[42] Mang Ye, Wei Shen, Junwu Zhang, Yao Yang, and Bo Du. 2024. Securereid: Privacy-preserving anonymization for person re-identification. *IEEE Transactions on Information Forensics and Security* (2024).

[43] Junwu Zhang, Mang Ye, and Yao Yang. 2022. Learnable privacy-preserving anonymization for pedestrian images. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7300–7308.

[44] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. 2017. Alignedreid: Surpassing human-level performance in person re-identification. arXiv 2017. *arXiv preprint arXiv:1711.08184* (2017).

[45] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 1116–1124.

[46] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*. 3754–3762.

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