Cross-video Identity Correlating for Person Re-identification Pre-training

Jialong Zuo [1] **Ying Nie** [2] **Hanyu Zhou** [1] **Huaxin Zhang** [1]
Haoyu Wang [2] **Tianyu Guo** [2] **Nong Sang** [1] **Changxin Gao** [1*]

[1] National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology,
[2] Huawei Noah's Ark Lab.
{jlongzuo, cgao}@hust.edu.cn

Abstract

Recent researches have proven that pre-training on large-scale person images extracted from internet videos is an effective way in learning better representations for person re-identification. However, these researches are mostly confined to pre-training at the instance-level or single-video tracklet-level. They ignore the identity-invariance in images of the same person across different videos, which is a key focus in person re-identification. To address this issue, we propose a Cross-video Identity-cOrrelating pre-traiNing (CION) framework. Defining a noise concept that comprehensively considers both intra-identity consistency and inter-identity discrimination, CION seeks the identity correlation from cross-video images by modeling it as a progressive multi-level denoising problem. Furthermore, an identity-guided self-distillation loss is proposed to implement better large-scale pre-training by mining the identity-invariance within person images. We conduct extensive experiments to verify the superiority of our CION in terms of efficiency and performance. CION achieves significantly leading performance with even fewer training samples. For example, compared with the previous state-of-the-art [9], CION with the same ResNet50-IBN achieves higher mAP of 93.3% and 74.3% on Market1501 and MSMT17, while only utilizing 8% training samples. Finally, with CION demonstrating superior model-agnostic ability, we contribute a model zoo named ReIDZoo to meet diverse research and application needs in this field. It contains a series of CION pre-trained models with spanning structures and parameters, totaling 32 models with 10 different structures, including GhostNet, ConvNext, RepViT, FastViT and so on.

1 Introduction

Person re-identification (ReID), which aims to identify and match a target person across different camera views, is becoming increasingly influential in widespread applications, such as criminal tracking and missing individual searching. Although current ReID methods [38, 2, 21, 22, 39] have achieved significant progress for the past few years, it has become evident through recent research advancements that the mere development of sophisticated specialist algorithms has already encountered a performance bottleneck.

Meanwhile, pre-training on large-scale images [1, 5, 17, 29] has shown great success to further improve model performance in the field of general vision. However, due to the significant domain gap, these pre-trained models can only provide very limited improvement for person re-identification. Therefore, with the vast amount of person-containing videos available on the internet, some researches [10, 11, 53, 9, 44, 27] turn to exploring the potential of pre-training on person images extracted from such videos and demonstrate encouraging results. However, due to the unavailability of identity labels for such extracted person images, these pre-training methods are confined to learning representations at the instance-level or single-video tracklet-level, as shown in Fig. 1

*Corresponding Author. Project Link: https://github.com/Zplusdragon/CION_ReIDZoo

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0795

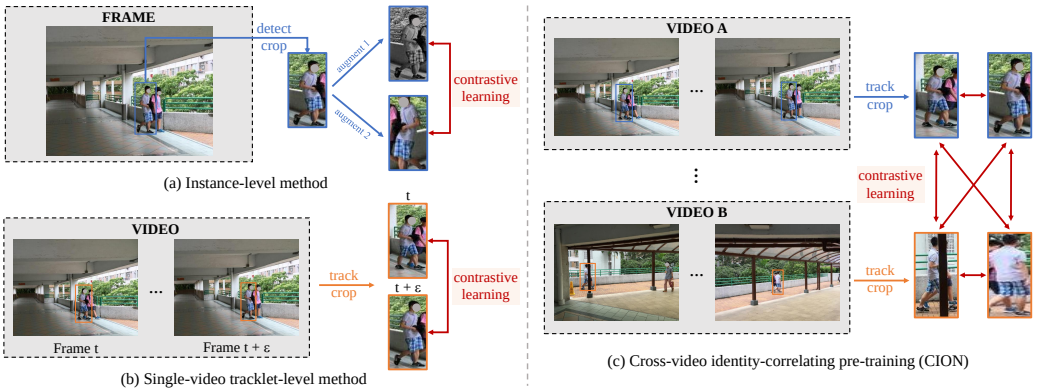

Figure 1: Comparisons between our proposed CION with other pre-training methods. In (a), the instance-level method mines *instance-invariance* by contrastive learning on augmented views of each image, completely ignoring the invariance within different images of the same person. In (b), the single-video tracklet-level method mines *tracklet-invariance* by contrastive learning on images of each tracklet in single video, significantly ignoring the invariance in images of the same person across different videos. In (c), our CION learns *identity-invariance* by correlating the images of the same person across different videos, thus leading to better representation learning.

On the one hand, the instance-level methods [10, 53, 44, 27] completely overlook the identity-invariance in different images of the same person. They incorporate some person-related prior knowledge into general pre-training frameworks, neglecting the concept of person identity. For example, LUP [10] improves MoCoV2 [4] by systematically studying the data augmentation and contrastive loss, while PASS [53] improves DINO [1] by generating part-level features to offer fine-grained information. However, there exists a notable concordance within different images of the same person. Directly ignoring such a concordance easily leads to learning weaker person representations.

On the other hand, the single-video tracklet-level methods [11, 9] only exploit the short-range tracklet-invariance while significantly neglecting the identity-invariance in images of the same person across different videos. They simply regard one tracklet in a single video as one person, which will result in two substantial problems. 1) Insufficient intra-identity consistency. Due to the unavoidable inaccurate tracking, there are quite a few images within a tracklet that do not belong to the same person. Roughly considering these images as coming from the same person will lead to insufficient intra-identity consistency. 2) Limited inter-identity discrimination. It is quite evident that there are considerable instances of the same person appearing in different videos, thus being treated as different tracklets (identities). Roughly considering these images as coming from different persons will lead to limited inter-identity discrimination.

To address these limitations mentioned above, as shown in Fig. 1, we propose a Cross-video Identity-cOrrelating pre-traiNing (CION) framework, by which the images of the same person across different videos are explicitly correlated to deeply mine the identity-invariance. CION first seeks the identity correlation from long-range cross-video images by modeling it as a progressive multi-level denoising problem, where a noise concept that comprehensively considers both intra-identity consistency and inter-identity discrimination is defined. Subsequently, with the sought identity correlation, CION introduces an identity-guided self-distillation loss to implement better pre-training by learning the invariance within augmented views in multiple images of the same person.

We conduct extensive experiments to verify the superiority of our CION in terms of efficiency and performance. CION achieves the best performance with significantly fewer training samples. With the ResNet50-IBN backbone, compared with the previous state-of-the-art method ISR [9], our CION achieves higher mAP of 93.3% and 74.3% on the Market1501 and MSMT17 datasets, while only utilizing 8% training samples. Noting that ISR is pre-training at the single-video tracklet-level, this experimental result fully validates the necessity of learning identity-invariance across different videos.

Finally, we contribute a fully open-sourced model zoo named ReIDZoo to meet diverse research and application needs within this field. With CION demonstrating superior performance improvements to diverse model structures (model-agnostic ability), we pre-trained a series of models with spanning structures and parameters, totaling 32 models with 10 different structures, including GhostNet [16], EdgeNext [28], ConvNext [26], RepViT [36], FastViT [35] and so on.

https://doi.org/10.52202/079017-0795

In conclusion, the highlights of our work can be summarized into three points:

(1) We propose a Cross-video Identity-cOrrelating pre-traiNing (CION) framework. It explicitly mines the identity-invariance in person images extracted from internet videos by progressive multi-level denoising and identity-guided self-distillation, leading to better representation learning.

(2) Extensive experiments verify the superiority of CION in terms of efficiency and performance. It achieves the best performance with much fewer training samples compared with previous methods, while also demonstrating adaptability to diverse model structures with spanning parameters.

(3) We contribute a model zoo named ReIDZoo. It contains a series of CION pre-trained models with spanning structures and parameters, totaling 32 models with 10 different structures. It conveniently meets various research and application needs, and greatly promotes the development of this field.

2 Related work

Self-supervised pre-training for general vision. In recent years, self-supervised pre-training methods have shown promising results in learning representation from large-scale data. MoCo [4] facilitates contrastive learning by building a dynamic dictionary with a queue and a moving-averaged encoder. BYOL [15] relies on two networks, referred to as online and target networks, that interact and learn from each other with a slow-moving average strategy. DINO [1] proposes a self-distillation framework by underlying the importance of momentum encoder, multi-crop training and so on. MAE [17] adopts the ViT backbone and learns representations by predicting the masked patches from the remaining visible ones. However, due to significant domain gap and neglect of person-related characteristics, these general methods show poor transferability to person re-identification.

Self-supervised pre-training for person re-identification. Existing pre-training methods can be classified into two categories, the first being the instance-level methods [10, 53, 44, 27, 3] and the second being single-video tracklet-level method [9, 11]. The instance-level methods strive to merge the person-related prior knowledge into existing self-supervised learning methods. For example, LUP [10] improves MoCov2 [4] through systematical studying of data augmentation and contrastive loss. PASS [53] improves DINO [1] by generating part-level features to offer fine-grained information. SOLIDER [3] turns to utilize pseudo semantic label and import some semantic information into the learned representations. However, these instance-level methods are incapable of harnessing the identity-invariance in different images of the same person. Meanwhile, the single-video tracklet-level methods strive to mine some invariance by regarding one tracklet in a single video as one person. For example, LUP-NL [10] utilizes supervised ReID learning, label-guided contrastive learning and prototype-based contrastive learning to rectify the noisy tracklet-level labels in person images extracted from raw internet videos. ISR [9] leverages massive data to construct the positive pairs from inter-frame images in a single videos. The identity-invarience of the same person across different videos is significantly neglected by these single-video tracklet-level methods.

3 CION: Cross-video Identity-cOrrelating pre-traiNing

We present a cross-video identity-correlating pre-training (CION) framework, which explicitly learns identity-invariance in long-range cross-video person images. First, we define a noise concept that comprehensively considers both intra-identity consistency and inter-identity discrimination (Sec. 3.1). Next, we model the identity correlation seeking process from cross-video images as a progressive multi-level denoising problem (Sec. 3.2). Finally, we propose an identity-guided self-distillation loss to mine identity-invariance from person images with the sought identity correlation (Sec. 3.3).

3.1 Noise definition

We demonstrate that a sample set with identity labels should have the following characteristics. 1) Intra-identity consistency. Samples belonging to the same identity should aggregate around a centroid in a high-dimensional space, i.e., they should be enveloped by a hypersphere $s_k = (\mathbf{c}_k, r_k)$, where $\mathbf{c}_k$ represents the centroid of the sphere and r_k represents its radius. A smaller radius of the sphere indicates a denser degree of aggregation, which signifies higher intra-identity consistency. 2) Inter-identity discrimination. Samples belonging to two different identities should be enveloped by two distinct hyperspheres s_i and s_j respectively, with a certain distance $d_{i,j}$ maintained between the

https://doi.org/10.52202/079017-0795

centroids of $\mathbf{c}_i$ and $\mathbf{c}_j$. A larger distance indicates a sparser distribution of hyperspheres, implying higher intra-identity discrimination. In summary, the smaller radii of the hyperspheres and the greater distances between the hypersphere centroids mean more desirable intra-identity consistency and inter-identity discrimination.

Considering the above demonstration, for a sample set $S = \{\mathbf{x}_i^{y_i}\}_{i=1}^{N}$, where the identity label $y_i \in \{1, 2, 3, \ldots, M\}$, we can further regard S as a hypersphere set $\{s_k\}_{k=1}^{M}$, where $\mathbf{x}_i \in s_k$ if $y_i = k$. Then for a single hypersphere s_k, the intra-consistency noise is formulated as:

$$n_k^{cst} = \mu(r_k - \sigma_{cst}) \quad s.t. \quad r_k = \max_i(dist(\mathbf{x}_i, \mathbf{c}_k)) \quad \forall \mathbf{x}_i \in s_k, \tag{1}$$

where $\mu(\cdot)$ is an unit step function, $dist(\cdot)$ is any distance metric such as cosine distance, and the centroid $\mathbf{c}_k$ is calculated by averaging all samples in s_k. We can observe that this formula determines the presence of noise by explicitly constraining the maximum radius of the hypersphere, indicating whether the aggregation degree of samples with the same identity meets the required criteria σ_{cst}.

Meanwhile, for two distinctive hyperspheres s_i and s_j, the inter-discrimination noise between them is formulated as:

$$n_{i,j}^{drm} = \mu(\sigma_{drm} - d_{i,j}) \quad s.t. \quad d_{i,j} = dist(\mathbf{c}_i, \mathbf{c}_j). \tag{2}$$

We can observe that this formula explicitly constrains the minimum distance between the centroids of the hyperspheres to determine the presence of noise, indicating whether the margin between samples of different identities meets the required criteria σ_{drm}.

Finally, comprehensively considering both intra-identity consistency and inter-identity discrimination, the overall noise for a sample set $S = \{s_k\}_{k=1}^{M}$ can be formulated as:

$$n_S = \frac{1}{M}(\sum_{k=1}^{M} n_k^{cst} + \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} n_{i,j}^{drm}). \tag{3}$$

We consider that a sample set with a favorable level of identity correlation should exhibit the smallest possible amount of n_S in order to achieve superior consistency and discrimination.

3.2 Progressive multi-level denoising

Given crawled internet videos, a straightforward approach [11, 9] is to utilize tracking algorithms to extract person tracklets from each video and treat each tracklet as an unique identity, thereby forming a sample set with initial identity correlation. However, due to inevitable false positives in tracking and the neglect of the same person across different videos, the identity correlation in this sample set is significantly noisy, lacking satisfactory intra-identity consisitency and inter-identity discrimination. Therefore, we propose a progressive multi-level denoising strategy to seek superior identity correlation by minimizing the overall n_S of the sample set as much as possible.

Single-tracklet denosing. First, we denoise each single tracklet to guarantee the intra-identity consistency. For an initial single-tracklet sample set with m extracted image features $t^0 = \{\mathbf{x}_i\}_{i=1}^{m}$, we obtain the denoised sample set by constraining the maximum hypersphere radius through multiple iterations. That is, for the sample set t^j of the j-th iteration, we compute the distances between each sample and the centroid of the remaining samples to identify the sample with the maximum deviation. If the deviation exceeds σ_{cst}, we remove the sample from t^j to an exclusion set t_{exc} and proceed to the next iteration; otherwise, we terminate the iteration and regard t^j as the denoised sample set with good intra-identity consistency. The iterative process can be formulated as:

$$\begin{cases} \mathbf{x}_{dev}^j = \arg\max_{\mathbf{x}_i} dist(\mathbf{x}_i, \mathbf{c}_{t^j \setminus \mathbf{x}_i}) \quad \forall \mathbf{x}_i \in t^j \\ t^{j+1} = t^j \setminus \mathbf{x}_{dev}^j, \quad t_{exc} = t_{exc} \cup \mathbf{x}_{dev}^j \quad s.t. \quad dist(\mathbf{x}_{dev}^j, \mathbf{c}_{t^j \setminus \mathbf{x}_{dev}^j}) \geq \sigma_{cst}, \end{cases} \tag{4}$$

We can observe that by iteratively constraining the maximum hypersphere radius of the tracklet, we achieve the exclusion of anomalous samples. In essence, this iterative process aims to eliminate the noise n^{cst} as defined in Eq. 1, thereby ensuring favorable intra-identity consistency for each tracklet.

Short-range single-video denosing. Then, we denoise each single video with multiple tracklets to guarantee both intra-identity consistency and inter-identity discrimination. For an initial single-video

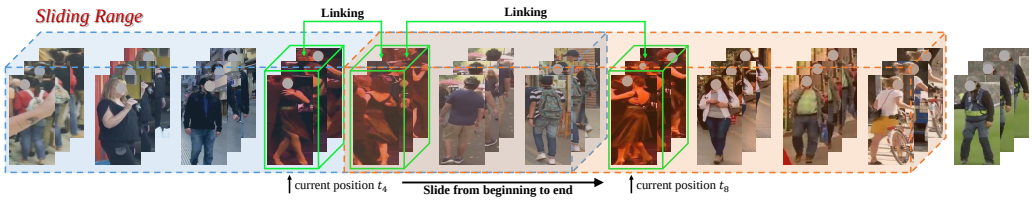

Figure 2: A toy example for Sliding Range and Linking Relation.

sample set with n denoised tracklets $v = \{t_k\}_{k=1}^{n}$and respective exclusion sets $\{t_{exc,k}\}_{k=1}^{n}$, we first reallocate the anomalous samples of each exclusion set to ensure their proper tracklet assignment. The reallocation process for an anomalous sample $\mathbf{x}_i^k \in t_{exc,k}$ can be formulated as:

$$\begin{cases} t_{min} = \arg\min_{t_r} dist(\mathbf{x}_i^k, \mathbf{c}_{t_r}) \quad \forall t_r \in v \backslash t_k \\ t_{min} = t_{min} \cup \mathbf{x}_i^k \quad s.t. \quad dist(\mathbf{x}_i^k, \mathbf{c}_{t_{min}}) < \sigma_{cst}. \end{cases} \tag{5}$$

Through the aforementioned reallocation process, we can assign each anomalous sample to its correct tracklet, or discard those extreme anomalous samples that do not belong to any tracklet, such as samples devoid of person presence. Subsequently, we merge tracklets with closely located hypersphere centroids to ensure robust inter-identity discrimination, which can be formulated as:

$$merge(t_p, t_q) = \mu(\sigma_{drm} - dist(\mathbf{c}_{t_p}, \mathbf{c}_{t_q})) \quad s.t. \quad \forall t_p, t_q \in v, \tag{6}$$

where $\mu(\cdot)$ is an unit step function to indicate whether the merge operation is applied to t_p and t_q.

We can observe that through the aforementioned two-stage process, we achieve the reallocation of anomalous samples and the merging of similar tracklets. In essence, this two-stage process aims to eliminate the overall noise n_S as defined in Eq. 3, thereby ensuring both favorable intra-identity consistency and inter-identity discrimination for each video.

Long-range cross-video denosing. Finally, we denoise multiple videos to seek identity correlation across videos. A straightforward method for denoising a large-scale sample set containing many videos is to concatenate the videos together and treat them as a single ultra-long video, then apply the single-video denoising strategy mentioned above. However, when a video contains N tracklets, the computational complexity of Eq. 6 is $\mathcal{O}(N^2)$. For an ultra-long video with an overwhelming number of tracklets, this undoubtedly introduces unmanageable computational overhead. To address this issue, we design Sliding Range and Linking Relation to implement ultra-long video denoising.

As shown in Fig. 2, for an ultra-long concatenated video composed of N sub-video tracklets $V = \{t_i\}_{i=1}^{N}$, we construct a Sliding Range $SR_c = \{c, r_s\}_{c:1 \to N}$ to slide from the beginning to the end of V, where c represents the index of the current position, and r_s represents the half-width of SR_c. Then, for all tracklets covered by SR_c, we obtain the Linking Relation between each tracklet and t_c by utilizing the merging function defined in Eq. 6, which is formulated as:

$$LR(t_i, t_c) = merge(t_i, t_c) \quad s.t. \quad \forall t_i \in SR_c \backslash t_c, \tag{7}$$

where $LR(\cdot)$ indicates whether a bidirectional Linking Relation will be established between the two tracklets. Subsequently, after the Sliding Range traverses the entire video, we perform a closure operation based on the obtained Linking Relations to derive the closures. Finally, the tracklets within each closure are merged as one single identity.

We can observe that the designing of Sliding Range significantly reduces the computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, which makes it possible to denoise massive-scale crawled videos. Meanwhile, the closure operation based on Linking Relations ensures long-range identity correlation for different tracklets of the same person across different videos. In essence, this long-range cross-video denosing strategy aims to eliminate the inter-discrimination noise n^{drm} as defined in Eq. 2 at the multi-video level, thereby further ensuring favorable inter-identity discrimination of all videos.

Identity correlation seeking. We summarize the whole process of seeking the identity correlation from large-scale crawled person-containing videos as follows. Firstly, we should utilize an off-the-shelf person tracking algorithm to extract person tracklets from each video and assign each tracklet with a class label, forming a noisy sample set with initial identity correlation. Then, the noisy sample set should undergo single-tracklet denoising, short-range single-video denoising and long-range cross-video denoising in sequence to seek satisfactory identity correlation. Finally, we consider all images within a single denoised tracklet as belonging to the same person, and this denoised sample set can serve as a well-prepared dataset with good identity correlation for subsequent pre-training.

https://doi.org/10.52202/079017-0795

3.3 Identity-guided self-distillation

With the sought identity correlation, we propose to utilize simple but effective identity-guided self-distillation to pre-train our models, leading to better identity-invariance learning. The whole framework shares a similar overall structure as the popular self-distillation pre-training paradigm DINO for general vision [1], as illustrated in Fig. 3. Differently, we introduce the concept of identity and conduct contrastive learning on the augmented views from images of the same person. We train a student network f_{θ_S} to match the output probability distribution of a teacher network f_{θ_T}, in which the two share the same architecture and are parameterized by θ_S and θ_T, respectively. Given an input image $\mathbf{x}$, both networks predict probability distributions over K dimensions denoted by P_S and P_T, which are obtained by normalizing the output of each network with a softmax function and formulated as:

$$P_S(\mathbf{x})^{(i)} = \frac{\exp\left(f_{\theta_S}(\mathbf{x})^{(i)}/\tau_s\right)}{\sum_{k=1}^{K}\exp\left(f_{\theta_S}(\mathbf{x})^{(k)}/\tau_s\right)}, \tag{8}$$

where $P_S(\mathbf{x})$ is the predicted distribution of the image $\mathbf{x}$ for f_{θ_S} and $\tau_s > 0$ is a hyperparameter controlling the output distribution sharpness. The similar formular holds for P_T with τ_t.

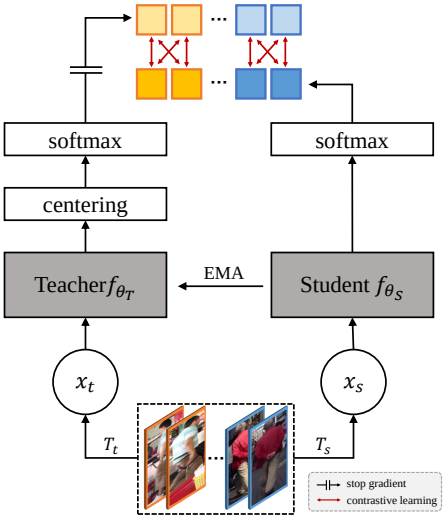

Figure 3: Self-distillation with identity guidance. The overall structure shares a similarity with [1], while the concept of identity is introduced. We illustrate it in the case of $N_{id} = 2$ and pairs of views $(\mathbf{x}_t, \mathbf{x}_s)$ for simplicity. T_t and T_s represent different random transformations. All transformed views from images of the same person will engage in contrastive learning.

In the following, we detail how we introduce the identity concept into self-distillation. First, with the sought identity correlation, we construct an identity set containing N_{id} images of the same person $X = \{\mathbf{x}_k\}_{k=1}^{N_{id}}$. Then, given an image $\mathbf{x}_k$, we construct a set V_k of different augmented views. This set contains two global views $\mathbf{x}_k^{g,1}$ and $\mathbf{x}_k^{g,2}$ and several local views of smaller resolution. We pass the whole set V_k through the student while only pass the global views through the teacher. Finally, the student learns to match the output probability distribution of teacher by minimizing the cross-entropy:

$$\min_{\theta_S} \sum_{\mathbf{x}_i \in X} \sum_{\mathbf{x}_j \in X} \sum_{\mathbf{x} \in \left\{\mathbf{x}_i^{g,1}, \mathbf{x}_i^{g,2}\right\}} \sum_{\substack{\mathbf{x}' \in V_j \\ \mathbf{x}' \neq \mathbf{x}}} H\left(P_T(\mathbf{x}), P_S\left(\mathbf{x}'\right)\right), \tag{9}$$

where $H(a, b) = -a \log b$ and the parameters θ_S are optimized with stochastic gradient descent. Following [1], the teacher network is frozen over an epoch and updated with an exponential moving average (EMA) on the student weights. Meanwhile, the output of the teacher network is centered with a mean calculated over the batch.

4 Experiments

4.1 Experimental setup

Pre-training and fine-tuning. During pre-training, we directly regard the images of SYNTHPEDES [55] as a noisy sample set with initial identity correlation. Then, we utilize our proposed progressive multi-level denoising strategy to seek the identity correlation in it. Meanwhile, to ensure a fair comparison with previous methods, we adopt random selection to align the overall sample size with that of the commonly used pre-training dataset LUPerson [10]. We name the whole sample set as CION-AL, and it contains 3,898,086 images of 246,904 identities totally. Finally, we utilize the CION-AL and our proposed identity-guided self-distillation loss to pre-train our models. We pre-trained 32 models of 10 architectures in total. During fine-tuning, as a common practice, we directly utilize TransReID [20] to fine-tune our pre-trained models with input size as 256×128 for the supervised person ReID setting, if not specified. Also, the commonly used C-Contrast [6] is adopted for the settings of unsupervised domain adaptation (UDA) and unsupervised learning (USL).

https://doi.org/10.52202/079017-0795

Datasets and evaluation metrics. Following common practices, we conduct experiments on two datasets, *i.e.*, Market1501 [50] and MSMT17 [40], for evaluating the performance on ReID tasks. Market1501 and MSMT17 are widely used person ReID datasets, which contain 32,688 images of 1,501 persons and 126,411 images of 4,101 persons, respectively. We use the cumulative matching characteristics at top-1 (Rank-1) and mean average precision (mAP) as the evaluation metrics.

Implementation details. To implement the progressive multi-level denoising, we utilize a ViT-Base model [27] trained on the images of UFine6926 [56] as a normalized feature extractor. The intra-consisitency criteria σ_{cst} and inter-discrimination criteria σ_{drm} are set as 0.2 and 0.18, respectively. We utilize cosine distance to calculate the distance between two samples. The half-width r_s of the sliding range is set as 1000. For pre-training, the global views and local views are resized to 256×128 and 128×64, respectively. We pre-train our models on $8\times$V100 GPUs for 100 epochs. We adopt a curriculum learning strategy for setting the images per identity, where the initial N_{id} is set to 2, and at epochs 40, 60, and 80, N_{id} increases to 4, 6, and 8, respectively. To optimize the utilization of GPU memory, the batch size per GPU and the number of cropped local views are adjusted according to the varying parameter counts of the models. All other settings for pre-training, such as learning rate, are consistent with those used in DINO [1]. For fine-tuning, we employ the commonly used methods [20, 38, 6] and make only slight adjustments to settings such as learning rate and batch size. Specific settings and more details can be found in Sec. A.5 of the appendix.

4.2 Comparison with state-of-the-art methods

Supervised person ReID. We compare our CION with several outstanding state-of-the-art methods on supervised person ReID in Tab. 1. We divide previous methods by model initialization, *i.e.*, utilizing supervised ImageNet pre-training [7] and self-supervised large-scale person images pre-training as backbone initialization, respectively. Compared with existing best methods pre-trained on ImageNet, our method significantly outperforms them without any extra complex designs. For example, our ViT-B achieves remarkable performance gains over the previous best method DCAL [52] on Market1501 and MSMT17 by 5.6% and 8.6% mAP, respectively. It should be noted that DCAL utilizes much more images for pre-training than ours (14M vs 3.9M). On the other hand, compared with existing best self-supervised methods pre-trained on large-scale person images, our method still shows significant superiority over them. It is worth mentioning that the single-video tracklet-level method ISR [9] utilizes a substantially larger number of person images for pre-training than ours (47.8M vs 3.9M). However, both fine-tuning with MGN [38], our ResNet50-IBN obtains 93.3% and 74.3% mAP on Market1501 and MSMT17, surpassing ISR by 1.0% and 2.8%, respectively. This effectively demonstrates the necessity of learning identity invariance across different videos.

Table 1: Comparison with SoTA methods of supervised person ReID. * means the result with the input size as 384×128. † represents fine-tuning with MGN.

Method	Backbone	Market1501		MSMT17	
		mAP	Rank1	mAP	Rank1
Pre-training on ImageNet1K-1.3M					
MGN* [38]	R50	87.5	95.1	63.7	85.1
ABDNet* [2]	R50	88.3	95.6	60.8	82.3
SAN [22]	R50	88.0	96.1	55.7	79.2
NFormer [39]	R50	91.1	94.7	59.8	77.3
Pre-training on ImageNet21K-14M					
TransReID [20]	ViT-B	87.4	94.6	63.6	82.5
DCAL [52]	ViT-B	87.5	94.7	64.0	83.1
Pre-training with Large-scale Person Images					
LUP†* (*4.2M*) [10]	R50	91.0	96.4	65.7	85.5
UPReID†* (*4.2M*) [44]	R50	91.1	97.1	63.3	84.3
LUP-NL†* (*10.7M*) [11]	R50	91.9	96.6	68.0	86.0
PLIP†* (*4.8M*) [55]	R50	91.3	96.7	66.2	85.3
ISR†* (*47.8M*) [9]	R50-IBN	92.3	96.9	71.5	88.4
Ours†* (*3.9M*)	R50	**92.3**	**97.1**	**70.1**	**87.8**
Ours†* (*3.9M*)	R50-IBN	**93.3**	**97.3**	**74.3**	**89.8**
MoCov3 (*4.2M*) [5]	ViT-S	82.2	92.1	47.4	70.3
TranSSL (*4.2M*) [27]	ViT-S	91.1	95.9	66.8	85.5
PASS (*4.2M*) [53]	ViT-S	92.2	96.3	69.1	86.5
PASS (*4.2M*) [53]	ViT-B	93.0	96.8	71.8	88.2
PATH (*11M*) [33]	ViT-B	89.5	95.8	69.1	84.3
Ours (*3.9M*)	ViT-S	**92.8**	**96.5**	**70.3**	**86.9**
Ours (*3.9M*)	ViT-B	**93.1**	**96.9**	**72.6**	**88.5**
SOLIDER* (*4.2M*) [3]	Swin-T	91.6	96.1	67.4	85.9
SOLIDER* (*4.2M*) [3]	Swin-S	93.3	96.6	76.9	90.8
Ours* (*3.9M*)	Swin-T	**92.2**	**96.8**	**71.1**	**88.1**
Ours* (*3.9M*)	Swin-S	**93.4**	**96.6**	**77.0**	**90.4**

Unsupervised person ReID. We compare our CION with some SoTA methods of unsupervised person ReID in Tab. 2, which contains two settings, *i.e.*, unsupervised domain adaptation (UDA) and unsupervised learning (USL). All the person-centric pre-trained models are fine-tuned with C-Contrast [6]. For both settings, our method achieves new SoTA performance, significantly surpassing all previous methods. For example, with ResNet50-IBN as backbone, our method outperforms the

https://doi.org/10.52202/079017-0795

Table 2: Comparison with state-of-the-art methods of unsupervised person ReID under two different settings. "Mar" and "MS" denote Market1501 and MSMT17 respectively. * denotes adding CFS [27].

(a) Comparison on UDA person ReID.

Method	Backbone	MS→Mar		Mar→MS	
		mAP	Rank1	mAP	Rank1
Pre-training on ImageNet1K-1.3M					
DG-Net++ [54]	R50	64.6	83.1	22.1	48.4
MMT [12]	R50	75.6	83.9	24.0	50.1
SpCL [13]	R50	77.5	89.7	26.8	53.7
C-Contrast [6]	R50	82.4	92.5	33.4	60.5
MCRN [41]	R50	-	-	32.8	64.4
Pre-training on Large-scale Person Images					
LUP [10]	R50	85.1	94.4	28.3	53.8
LUP [10]	R50-IBN	86.9	84.6	42.6	69.1
TranSSL [27]	ViT-S	89.6	95.6	55.0	77.9
PASS [53]	ViT-S	90.2	95.8	49.1	72.7
TranSSL* [27]	ViT-S	89.9	95.5	57.8	79.5
Ours	R50	**90.6**	**96.1**	**57.4**	**81.3**
Ours	R50-IBN	**92.1**	**96.5**	**62.7**	**84.3**
Ours	ViT-S	**90.4**	**95.6**	**58.0**	**79.9**

(b) Comparison on USL person ReID.

Method	Backbone	Mar		MS	
		mAP	Rank1	mAP	Rank1
Pre-training on ImageNet1K-1.3M					
MMCL [37]	R50	45.5	80.3	11.2	35.4
HCT [49]	R50	56.4	80.0	-	-
IICS [43]	R50	72.9	89.5	26.9	52.4
C-Contrast [6]	R50	82.6	93.0	33.1	63.3
MCRN [41]	R50	80.8	92.5	31.2	63.6
Pre-training on Large-scale Person Images					
LUP [10]	R50	84.0	93.4	31.4	58.8
LUP [10]	R50-IBN	86.4	94.2	39.8	66.1
TranSSL [27]	ViT-S	89.3	94.8	48.8	74.4
PASS [53]	ViT-S	88.5	94.9	41.0	67.0
TranSSL* [27]	ViT-S	89.6	95.3	50.6	75.0
Ours	R50	**90.6**	**95.7**	**52.5**	**78.3**
Ours	R50-IBN	**91.4**	**96.1**	**59.4**	**82.4**
Ours	ViT-S	**90.6**	**95.6**	**55.0**	**78.4**

previous SoTA LUP [27] by a large margin, *i.e.*, 20.1% mAP on Market1501 → MSMT17 UDA setting and 19.6% mAP on MSMT17 USL setting, respectively. Noting that LUP is an instance-level method, these results verify the superiority of our identity-level method in person ReID pre-training.

4.3 Generalizing to different model structures

During pre-training, CION imposes minimal constraints on the model structure, allowing for its generalization across a variety of structures. To verify this, we conduct extensive experiments to investigate whether CION can improve the models with different structures and parameters, encompassing 32 models with 10 structures, including CNNs, ViTs, and other specialized architectures. The baselines are the models with supervised ImageNet [7] pre-training. As a common practice, we utilize TransReID [20] to fine-tune the models on ReID datasets with input size as 256×128, if not specified. As the results listed in Tab. 3, our CION consistently yields significant performance improvements across all models. When applying to TransReID, the average performance improvements of all the 32 models over the baselines are 11.7% and 18.8% mAP on Market1501 and MSMT17, respectively. With CION demonstrating such model-agnostic ability, we construct a model zoo named ReIDZoo, which contains all the CION pre-trained models with spanning structures and parameters, to meet diverse research and application needs in this field.

Table 3: CION significantly improves different models. FLOPs are computed based on the input size. * means the result with the input size as 384×128. † represents fine-tuning with MGN. ↑ indicates the improvement brought by our CION over the baseline.

Model	Params. (M)	FLOPs (G)	Market1501		MSMT17	
			mAP	Rank1	mAP	Rank1
ResNet50†* [18]	23.51	6.14	92.3(↑**4.7**)	97.1(↑**2.1**)	70.1(↑**6.6**)	87.8(↑**2.8**)
ResNet101†* [18]	42.50	9.80	93.1(↑**4.9**)	97.1(↑**1.8**)	71.7(↑**6.9**)	88.8(↑**2.9**)
ResNet152†* [18]	58.14	13.47	93.1(↑**4.4**)	97.3(↑**1.6**)	72.4(↑**6.5**)	89.2(↑**3.1**)
ResNet50-IBN†* [42]	23.51	6.14	93.3(↑**4.1**)	97.3(↑**1.5**)	74.3(↑**8.9**)	89.8(↑**3.1**)
ResNet101-IBN†* [42]	42.50	9.80	93.8(↑**4.3**)	97.2(↑**1.3**)	76.0(↑**8.4**)	90.7(↑**3.1**)
ResNet152-IBN†* [42]	58.14	13.47	94.3(↑**4.5**)	97.3(↑**1.6**)	76.8(↑**8.6**)	90.8(↑**3.3**)
GhostNet 0.5× [16]	1.31	0.03	81.0(↑**10.7**)	92.0(↑**5.8**)	39.6(↑**10.6**)	65.4(↑**9.1**)
GhostNet 1.0× [16]	3.90	0.10	87.7(↑**11.5**)	95.1(↑**6.1**)	53.7(↑**12.0**)	76.9(↑**9.8**)
GhostNet 1.3× [16]	6.08	0.16	88.8(↑**12.4**)	95.2(↑**5.9**)	55.4(↑**11.7**)	78.1(↑**9.4**)
EdgeNext-XS [28]	2.14	0.27	88.7(↑**15.9**)	94.7(↑**6.5**)	57.9(↑**18.1**)	79.8(↑**13.7**)
EdgeNext-S [28]	5.28	0.63	91.3(↑**14.5**)	96.1(↑**5.9**)	63.2(↑**19.8**)	83.2(↑**13.7**)
EdgeNext-B [28]	17.93	1.92	92.4(↑**14.7**)	96.8(↑**6.5**)	68.6(↑**22.3**)	86.6(↑**14.7**)
RepViT-M0.9 [36]	4.72	0.56	92.8(↑**12.8**)	96.9(↑**5.8**)	68.4(↑**19.8**)	86.5(↑**12.0**)
RepViT-M1.0 [36]	6.40	0.75	93.1(↑**25.6**)	96.6(↑**11.4**)	69.5(↑**47.5**)	87.2(↑**43.4**)
RepViT-M1.5 [36]	13.62	1.54	92.8(↑**23.2**)	96.6(↑**10.3**)	69.5(↑**35.1**)	87.5(↑**26.2**)
FastViT-S12 [35]	8.45	0.93	91.2(↑**12.3**)	96.4(↑**5.0**)	59.0(↑**17.3**)	81.4(↑**13.6**)
FastViT-SA12 [35]	10.56	1.00	90.9(↑**14.7**)	95.8(↑**5.9**)	59.0(↑**17.1**)	81.6(↑**13.8**)
FastViT-SA24 [35]	20.53	1.92	92.9(↑**19.9**)	96.9(↑**8.7**)	69.6(↑**32.6**)	87.6(↑**23.8**)
ResNet18 [18]	11.18	2.00	89.3(↑**10.9**)	95.6(↑**4.5**)	55.6(↑**14.8**)	79.6(↑**12.0**)
ResNet50 [18]	23.51	4.09	92.8(↑**9.9**)	96.8(↑**3.9**)	67.7(↑**15.5**)	86.4(↑**9.8**)
ResNet101 [18]	42.50	6.54	94.0(↑**10.8**)	97.3(↑**4.6**)	73.2(↑**17.3**)	88.8(↑**9.0**)
ResNet152 [18]	58.14	8.98	94.2(↑**8.1**)	97.1(↑**3.0**)	74.5(↑**18.2**)	89.3(↑**10.4**)
ResNet18-IBN [42]	11.18	2.00	90.0(↑**10.2**)	96.0(↑**4.4**)	56.1(↑**13.5**)	79.5(↑**10.4**)
ResNet50-IBN [42]	23.51	4.09	93.7(↑**8.9**)	96.9(↑**3.1**)	72.3(↑**18.6**)	88.5(↑**11.0**)
ResNet101-IBN [42]	42.50	6.54	94.3(↑**7.9**)	97.3(↑**3.1**)	75.8(↑**18.0**)	89.8(↑**9.6**)
ResNet152-IBN [42]	58.14	8.98	94.6(↑**7.2**)	97.6(↑**3.0**)	76.7(↑**17.8**)	90.3(↑**8.9**)
ConvNext-T [26]	27.82	2.92	92.0(↑**14.1**)	96.9(↑**6.0**)	68.9(↑**22.7**)	87.3(↑**14.8**)
ConvNext-S [26]	49.45	5.68	92.9(↑**12.5**)	96.8(↑**5.2**)	69.9(↑**21.8**)	88.0(↑**14.4**)
ConvNext-B [26]	87.57	10.04	93.4(↑**13.2**)	96.9(↑**5.3**)	72.7(↑**24.1**)	89.0(↑**14.5**)
VOLO-D1 [47]	25.84	4.40	93.0(↑**7.8**)	96.9(↑**3.3**)	73.9(↑**15.8**)	88.7(↑**8.6**)
VOLO-D2 [47]	57.62	9.15	92.1(↑**7.1**)	96.4(↑**2.9**)	70.9(↑**11.6**)	87.5(↑**6.6**)
VOLO-D3 [47]	85.26	13.26	90.5(↑**5.6**)	95.7(↑**1.8**)	62.8(↑**2.1**)	83.0(↑**1.3**)
ViT-T [8]	6.43	1.55	91.4(↑**10.5**)	96.3(↑**5.1**)	65.0(↑**18.9**)	84.6(↑**14.9**)
ViT-S [8]	23.39	3.79	92.8(↑**8.0**)	96.5(↑**2.6**)	70.3(↑**16.7**)	86.9(↑**11.6**)
ViT-B [8]	89.15	12.37	93.1(↑**6.4**)	96.9(↑**2.3**)	72.6(↑**11.4**)	88.5(↑**6.6**)
Swin-T [25]	27.53	3.13	92.5(↑**9.8**)	97.0(↑**4.3**)	71.1(↑**19.6**)	87.6(↑**12.6**)
Swin-S [25]	48.85	5.93	93.6(↑**9.1**)	96.8(↑**3.6**)	76.0(↑**19.7**)	89.6(↑**10.9**)
Swin-B [25]	86.76	10.44	94.0(↑**9.6**)	96.9(↑**3.0**)	76.7(↑**20.7**)	90.3(↑**12.1**)
Average Impr.			↑**11.7**	↑**5.0**	↑**18.8**	↑**12.9**

https://doi.org/10.52202/079017-0795

4.4 Ablation studies and analyses

Identity-level correlating holds significance. We validate the significance of identity-level correlating in learning better person representations through both feature distribution visualization and performance comparison. For visualization, we collect samples of 15 persons randomly selected from the gallery of Market1501 [50], and visualize their features via t-SNE [34] in Fig. 4. The instance-level method LUP [10] and the tracklet-level method LUP-NL [11], which are respectively pre-trained on 4.2M and 10.7M person images, are employed as a comparison. As we can see, the features of LUP are almost randomly distributed, while the features of LUP-NL remain challenging to aggregate at the identity level. However, our CION consistently enjoys good intra-consistency and inter-discrimination.

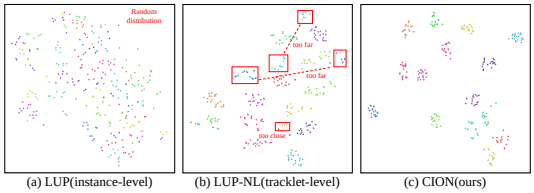

Figure 4: The t-SNE [34] visualization of extracted image features. Our CION enjoys good identity consistency and discrimination, while the instance-level method LUP [10] and tracklet-level method LUP-NL [11] do not (marked in red).

Table 4: Our identity-level method achieves significant leading performance compared with two identity-ignored pre-training methods [10, 1].

Method	Pre-train	Market1501		MSMT17	
		mAP	Rank1	mAP	Rank1
LUP+MGN	CION-AL	91.4	96.6	68.2	86.4
DINO+MGN	CION-AL	90.9	96.3	66.7	85.8
Ours+MGN	CION-AL	**93.3**	**97.3**	**74.3**	**89.8**

For performance comparison, we compare the fine-tuning results of our CION with two popular identity-ignored pre-training methods LUP [10] and DINO [1] in Tab. 4. For a fair comparison, the models are all ResNet50-IBN [42] pre-trained on our CION-AL and the fine-tuning algorithm is MGN [38]. We can see that the model pre-trained with CION achieves significant leading performance. These results demonstrate the significance of implementing identity-level correlating for person ReID pre-training.

Effectiveness of progressive multi-level denoising. Fig. 5 shows the ablation study of each denoising strategy. The number of training samples for all groups is set to 3.9M for a fair comparison. The backbone is ResNet50-IBN [42] and the fine-tuning algorithm is MGN [38]. We can observe that each strategy significantly improves the fine-tuning performance of the pre-trained model on downstream datasets. Particularly, the cross-video denoising results in the most substantial improvement, further increasing the mAP by 0.6% and 1.9% on Market1501 and MSMT17, respectively. These results verify the effectiveness of our progressive multi-level denoising strategy, which guarantees satisfactory identity correlation seeking.

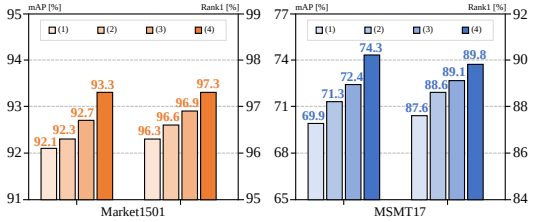

Figure 5: Results with different denoising strategies: (1) without denoising, (2) add single-tracklet denoising, (3) further add single-video denoising, (4) further add cross-video denoising.

Scalability to large-scale training data. We also study the scalability of our method to large-scale training data. Aside from the training data size, the rest settings are same with the experiment mentioned above. As shown in Fig 6, with training data size ratio varying from 10% to 100%, the fine-tuning performance on downstream datasets consistently keeps improving, especially on the more challenging MSMT17. This indicates our method's potential for further improved performance with increasing training data.

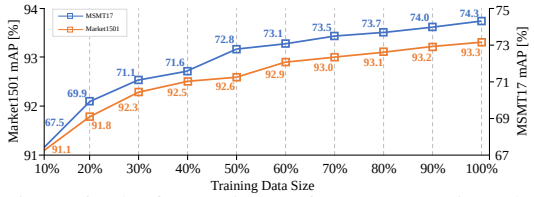

Figure 6: The fine-tuning performance consistently keeps improving as the training data increases.

5 Conclusion

We present a novel framework, CION, to learn identity-invariance from cross-video person images for person re-identification pre-training. In particular, we model the identity correlation seeking process as a progressive multi-level denoising problem, with a novel noise concept defined. Also, we propose an identity-guided self-distillation loss to implement better large-scale pre-training. Compared to existing pre-training methods, we achieve significantly leading performance with even fewer training samples. Meanwhile , we contribute a model zoo to promote the development of this field.

https://doi.org/10.52202/079017-0795

Acknowledgements. This work was supported by the National Natural Science Foundation of China No.62176097, and the Hubei Provincial Natural Science Foundation of China No.2022CFA055. We gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021.

[2] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, pages 8351–8361, 2019.

[3] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *CVPR*, pages 15050–15061, 2023.

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021.

[6] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *ACCV*, pages 1142–1160, 2022.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[9] Zhaopeng Dou, Zhongdao Wang, Yali Li, and Shengjin Wang. Identity-seeking self-supervised representation learning for generalizable person re-identification. In *ICCV*, pages 15847–15858, 2023.

[10] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *CVPR*, pages 14750–14759, 2021.

[11] Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. Large-scale pre-training for person re-identification with noisy labels. In *CVPR*, pages 2476–2486, 2022.

[12] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020.

[13] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *NeurIPS*, 33:11309–11321, 2020.

[14] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4313–4322, 2022.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020.

[16] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *CVPR*, pages 1580–1589, 2020.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[19] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020.

[20] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021.

[21] Jiahao Hong, Jialong Zuo, Chuchu Han, Ruochen Zheng, Ming Tian, Changxin Gao, and Nong Sang. Spatial cascaded clustering and weighted memory for unsupervised person re-identification. *arXiv preprint arXiv:2403.00261*, 2024.

[22] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, volume 34, pages 11173–11180, 2020.

[23] He Li, Mang Ye, and Bo Du. Weperson: Learning a generalized re-identification model from all-weather virtual data. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3115–3123, 2021.

[24] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17459–17469, 2024.

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CVPR*, 2022.

[27] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*, 2021.

[28] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *ECCV*, pages 3–20. Springer, 2022.

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

[30] Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *Computer Vision–ECCV 2024: 18th European Conference*, page 456–474, 2024.

[31] Jiangming Shi, Xiangbo Yin, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. *arXiv preprint arXiv:2402.19026v2*, 2024.

[32] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11218–11228, 2023.

[33] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *CVPR*, pages 21970–21982, 2023.

[34] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The journal of machine learning research*, 15(1):3221–3245, 2014.

[35] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *ICCV*, pages 5785–5795, 2023.

[36] Ao Wang, Hui Chen, Zijia Lin, Hengjun Pu, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. *CVPR*, 2024.

[37] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, pages 10981–10990, 2020.

 https://doi.org/10.52202/079017-0795

[38] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018.

[39] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. Nformer: Robust person re-identification with neighbor transformer. In *CVPR*, pages 7297–7307, 2022.

[40] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[41] Yuhang Wu, Tengteng Huang, Haotian Yao, Chi Zhang, Yuanjie Shao, Chuchu Han, Changxin Gao, and Nong Sang. Multi-centroid representation network for domain adaptive person re-id. In *AAAI*, volume 36, pages 2750–2758, 2022.

[42] Pan Xingang, Luo Ping, Shi Jianping, and Tang Xiaoou. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018.

[43] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *CVPR*, pages 11926–11935, 2021.

[44] Zizheng Yang, Xin Jin, Kecheng Zheng, and Feng Zhao. Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification. In *CVPR*, pages 14298–14307, 2022.

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

[46] Xiangbo Yin, Jiangming Shi, Yachao Zhang, Yang Lu, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Robust pseudo-label learning with neighbor relation for unsupervised visible-infrared person re-identification. *arXiv preprint arXiv:2405.05613*, 2024.

[47] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE TPAMI*, 45(5):6575–6586, 2022.

[48] Gong Yunpeng, Zhong Zhun, Luo Zhiming, Qu Yansong, Ji Rongrong, and Jiang Min. Cross-modality perturbation synergy attack for person re-identification. *arXiv preprint arXiv:2401.10090*, 2024.

[49] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *CVPR*, pages 13657–13665, 2020.

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[51] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017.

[52] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *CVPR*, pages 4692–4702, 2022.

[53] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *ECCV*, pages 198–214. Springer, 2022.

[54] Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Vijaya Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*, pages 87–104. Springer, 2020.

[55] Jialong Zuo, Jiahao Hong, Feng Zhang, Changqian Yu, Hanyu Zhou, Changxin Gao, Nong Sang, and Jingdong Wang. Plip: Language-image pre-training for person representation learning, 2024.

[56] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. *CVPR*, 2024.

A Appendix

A.1 Broader impact

This paper proposes a cross-video identity-correlating pre-training framework for person re-identification, and contributes a pre-trained model zoo with spanning model structures and parameters. Our framework and models can help to identity and track pedestrians across different cameras, thus boosting the development of smart retail, smart transportation, smart security systems and so on in the future metropolises. In addition, our proposed pre-training framework is quite general and not limited to the specific research field of person ReID. It can be well extended to broader research areas with identity concept such as vehicle ReID.

Nevertheless, the application of person ReID, such as for identifying and tracking pedestrians in surveillance systems [14, 48, 31, 32, 30], might raise privacy concerns. It typically depends on the utilization of surveillance data for training without the explicit consent of the individuals being recorded. Therefore, governments and officials need to carefully establish strict regulations and laws to control the usage of ReID technologies. Otherwise, person ReID technologies [46, 24, 23, 45] can potentially equip malicious actors with the ability to surveil pedestrians through multiple cameras without their consent. Meanwhile, the research community should also avoid using any dataset with ethics issues. For example, DukeMTMC [51] has been taken down due to the violation of data collection terms and should no longer be used. We would not evaluate our models on DukeMTMC related benchmarks as well. Furthermore, we should be cautious of the misidentification of the ReID systems to avoid possible disturbance. Also, note that the demographic makeup of the datasets used is not representative of the broader population.

At the same time, we have utilized a substantial amount of person-containing video data from the internet for pre-training purposes. Consequently, it is inevitable that the resulting models may inherently contain information about these persons. Researchers should adhere to relevant laws and regulations, and strive to avoid using our models for any improper invasion of privacy. We will have a gated release of our models and training data to avoid any misuse. We will require that users adhere to usage guidelines and restrictions to access our models and training data. Meanwhile, all our open-sourced assets can only be used for research purpose and are forbidden for any commercial use.

A.2 Limitations

CION presents a preliminary attempt to correlate person images across different internet videos to implement person re-identification pre-training. Despite its effectiveness on existing public datasets, CION may still be difficult to learn good fine-grained person representations for it does not explicitly achieve further fine-grained information mining. Meanwhile, due to the limitations of the feature extractor's representational capacity, it is inevitable that some extremely difficult noise will be hard to remove during the denoising process. As a result, the automatically sought identity correlation may still have a certain discrepancy in accuracy when compared to manually annotated correlation, which may introduce a certain degree of ambiguity during the training process. Additionally, the Sliding Range and Linking Relation strategy we proposed to conserve computational resources may result in difficulty seeking correlations between tracklets of the same person across different videos that are spaced extremely far apart. Also, as we have followed the conventional practice of previous works by utilizing a large amount of person-containing internet video data to implement pre-training, there is a potential for privacy and security issues to some extent. Therefore, in our subsequent work, we will focus on addressing fine-grained issues and improving the accuracy of identity correlations. We will take every possible measure to prevent the misuse of our models and dataset as well.

A.3 Safeguards for dataset and models

To address the privacy concerns associated with crawling videos from the internet and the application of person ReID technology, we will implement a controlled release of our models and dataset, thereby preventing privacy violations and ensuring information security. We require that the crawled images are sourced from legitimate video websites to avoid the inclusion of harmful content. We will also require that users adhere to usage guidelines and restrictions to access our models and dataset. For instance, we have drafted the following regulations that must be adhered to, which will be refined and elaborated in subsequent releases:

https://doi.org/10.52202/079017-0795

1. Privacy: All individuals using the models in ReIDZoo and the CION-AL dataset should agree to protect the privacy of all the subjects in it. The users should bear all responsibilities and consequences for any loss caused by the misuse.

2. Redistribution: The models in ReIDZoo and the CION-AL dataset, either entirely or partly, should not be further distributed, published, copied, or disseminated in any way or form without a prior approval from the creators, no matter for profitable use or not.

3. Commercial Use: The models in ReIDZoo and the CION-AL dataset, entirely, partly, or in any format derived thereof, is not allowed for commercial use.

4. Modification: The models in ReIDZoo and the CION-AL dataset, either entirely or partly, is not allowed to be modified.

In parallel, we will require users to provide relevant information and will rigorously screen the submitted details to restrict access to our dataset and models by institutions or individuals with a history of privacy violations.

A.4 Licenses for existing assets

We utilize some existing assets to conduct our research, which can be categorized to three types: code, data and models. All assets used in our paper are properly credited and all original paper are cited. We list the license and URL for each asset below.

Code

1. Emerging Properties in Self-Supervised Vision Transformers [1]. (DINO)

Apache 2.0 License, URL: https://github.com/facebookresearch/dino

2. TransReID: Transformer-based Object Re-Identification [20]. (TransReID)

MIT License, URL: https://github.com/damo-cv/TransReID

3. Cluster Contrast for Unsupervised Person Re-Identification [6]. (C-Contrast)

MIT License, URL: https://github.com/alibaba/cluster-contrast-reid

4. Unsupervised Pre-training for Person Re-identification [10]. (LUP)

MIT License, URL: https://github.com/DengpanFu/LUPerson

5. Self-Supervised Pre-Training for Transformer-Based Person Re-Identification [27]. (TranSSL)

MIT License, URL: https://github.com/damo-cv/TransReID-SSL

6. FastReID: A Pytorch Toolbox for General Instance Re-identification [19]. (FastReID)

Apache 2.0 License, URL: https://github.com/JDAI-CV/fast-reid

Data

1. PLIP: Language-Image Pre-training for Person Representation Learning [55]. (SYNTH-PEDES)

MIT License, URL: https://github.com/Zplusdragon/PLIP

2. UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity [56]. (UFine6926)

UFineBench License, URL: https://github.com/Zplusdragon/UFineBench

3. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification [40]. (MSMT17)

MSMT17 Release Agreement, URL: https://www.pkuvmc.com/dataset.html

4. Scalable Person Re-Identification: A Benchmark [50]. (Market1501)

MIT License, URL: http://zheng-lab.cecs.anu.edu.au/Project/project_reid.html

Models

1. Deep Residual Learning for Image Recognition [18]. (ResNet)

BSD-3-Clause license, URL: https://github.com/pytorch/vision

2. Instance-Batch Normalization Network [42]. (ResNet-IBN)

MIT License, URL: https://github.com/XingangPan/IBN-Net

3. GhostNet: More Features from Cheap Operations [16]. (GhostNet)

Apache 2.0 License, URL: https://github.com/huawei-noah/Efficient-AI-Backbones/

4. EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications [28]. (EdgeNext)

MIT License: URL: https://github.com/mmaaz60/EdgeNeXt

5. RepViT: Revisiting Mobile CNN From ViT Perspective [36]. (RepViT)

Apache 2.0 License, URL: https://github.com/THU-MIG/RepViT

6. FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization [35]. (FastViT)

FastViT License, URL: https://github.com/apple/ml-fastvit

7. A ConvNet for the 2020s [26]. (ConvNext)

MIT License, URL: https://github.com/facebookresearch/ConvNeXt

8. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (ViT)

Apache 2.0 License, URL: https://github.com/google-research/vision_transformer

9. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows [25]. (Swin)

MIT License, URL: https://github.com/microsoft/Swin-Transformer

10. VOLO: Vision Outlooker for Visual Recognition [47]. (VOLO)

Apache 2.0 License, URL: https://github.com/sail-sg/volo

A.5 Pre-training setting for each model

Due to the varying memory requirements of different models during the pre-training process, we optimize the utilization of computational resources by setting different batch sizes and numbers of local cropped views for each model. We report the batch sizes, numbers of local cropped views and rough training time for each model in Tab. 5. Each model is trained on 8 $\times$ V100 GPUs. The pre-training dataset is the whole CION-AL, which has 3,898,086 images of 246,904 persons.

Table 5: The batch sizes, numbers of local cropped views and rough training time for each model.

Model	batch size	local views	time(d≈)
GhostNet 0.5× [16]	216	8	4.8
GhostNet 1.0× [16]	192	8	5.7
GhostNet 1.3× [16]	168	8	7.2
EdgeNext-XS [28]	192	8	5.1
EdgeNext-S [28]	144	8	6.5
EdgeNext-B [28]	120	8	7.3
RepViT-M0.9 [36]	144	8	5.1
RepViT-M1.0 [36]	144	8	6.2
RepViT-M1.5 [36]	96	8	8.1
FastViT-S12 [35]	120	8	5.3
FastViT-SA12 [35]	120	8	6.8
FastViT-SA24 [35]	96	8	9.3
ResNet18 [18]	192	10	3.6
ResNet50 [18]	120	8	5.1
ResNet101 [18]	96	8	6.6
ResNet152 [18]	72	8	9.7
ResNet18-IBN [42]	192	10	3.6
ResNet50-IBN [42]	120	8	5.1
ResNet101-IBN [42]	96	8	6.6
ResNet152-IBN [42]	72	8	9.7
ConvNext-T [26]	120	7	6.2
ConvNext-S [26]	96	6	7.3
ConvNext-B [26]	72	6	9.5
ViT-T [8]	120	8	5.3
ViT-S [8]	96	8	6.7
ViT-B [8]	72	8	8.6
Swin-T [25]	96	7	5.8
Swin-S [25]	72	6	8.3
Swin-B [25]	48	8	11.6
VOLO-D1 [47]	120	4	6.2
VOLO-D2 [47]	72	4	9.4
VOLO-D3 [47]	48	4	13.3

A.6 Samples used in t-SNE visualization

For reproducing the t-SNE visualization results shown in Fig. 4, we display the samples we randomly selected from Market1501 [50] in Fig. 8. There are totally 15 persons with 20 cropped images each. We obscure the faces of each person to avoid the privacy violations. Researchers can use these samples and the official models of LUP [10] and LUP-NL [11] to reproduce the visualization results.

A.7 Linear training time increasing

We have assessed how the training time of CION scale with pre-training dataset size. We use ResNet50 to conduct the experiment. We record

https://doi.org/10.52202/079017-0795

the training time required by CION as the pre-training data size increases from 10% to 100%. As shown in Fig. 7, the training time of our CION increases almost linearly with the growth in the size of the pre-trained dataset. This linear increase in training time ensures that our CION can be effectively applied to large-scale pre-training. This indicates that we can achieve better performance by adding more training data without incurring an unbearable training cost.

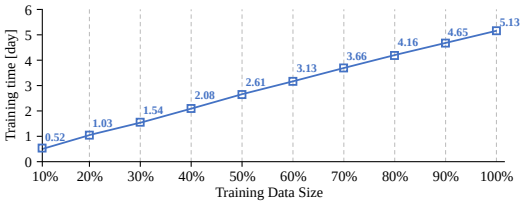

Figure 7: The time consumption for pre-training on CION-AL with different data size ratio.

A.8 More results for unsupervised person ReID

To further assess the model structure adaptability of our CION, we evaluate the performance of more CION pre-trained models for unsupervised person ReID, inclucing ResNet [18], ResNet-IBN [42] and ViT [8]. We report the results in Tab. 6. As we can see, all the models achieve superior performance for different unsupervised settings on Market1501 [50] and MSMT17 [40]. Meanwhile, as the number of parameters increases, the performance of each model consistently improves. These results further indicate that our CION can be effectively used to pre-train various models, enabling them to achieve significant performance improvements for person ReID.

Table 6: The performance of different CION pre-trained models for unsupervised person ReID. "Mar" and "MS" denote Market1501 and MSMT17 respectively.

Model	MS→Mar		Mar→MS		Market1501		MSMT17	
	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1
ResNet50 [18]	90.6	96.1	57.4	81.3	90.6	95.7	52.5	78.3
ResNet101 [18]	93.1	96.8	67.6	86.9	93.2	97.0	64.2	85.2
ResNet152 [18]	93.6	97.0	70.1	88.1	93.3	97.1	68.8	87.6
ResNet50-IBN [42]	92.1	96.5	62.7	84.3	91.4	96.1	59.4	82.4
ResNet101-IBN [42]	93.7	96.8	70.9	88.2	93.8	96.7	69.7	88.2
ResNet152-IBN [42]	94.0	97.1	73.7	89.2	94.3	97.1	71.7	89.0
ViT-T [8]	88.0	94.7	44.3	69.3	88.3	94.5	43.1	69.6
ViT-S [8]	90.4	95.6	58.0	79.9	90.6	95.6	55.0	78.4
ViT-B [8]	91.1	95.9	61.4	82.9	91.2	96.1	58.9	82.3

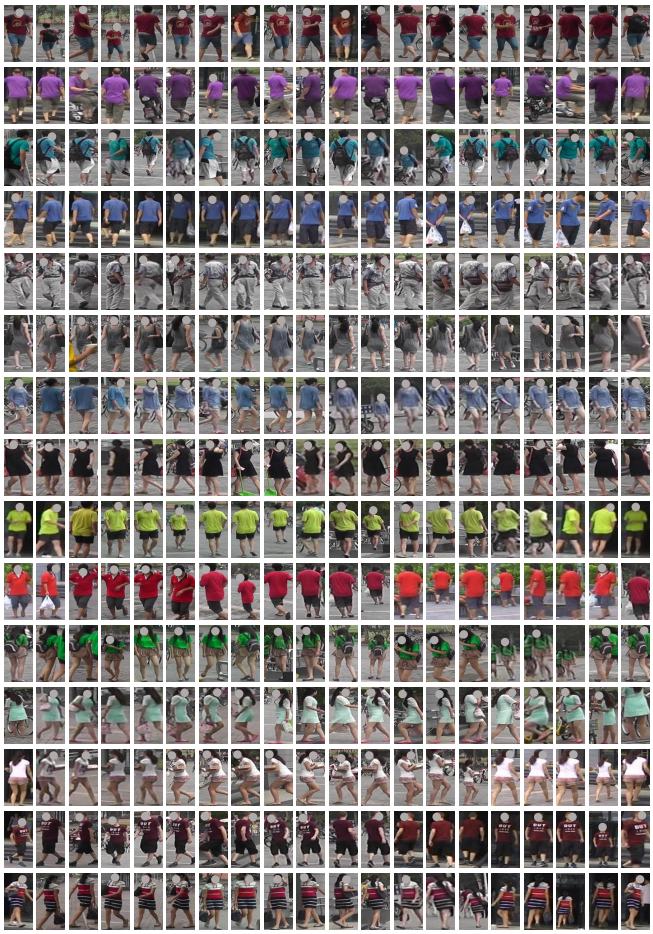

Figure 8: The samples used in t-SNE feature distribution visualization.

https://doi.org/10.52202/079017-0795

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction of our paper accurately reflect our paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We conduct an in-depth discussion of the limitations of our work, which can be found in Sec. A.2 of the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-0795

Answer: [NA]

Justification: Our work focuses on computer vision applications, not including theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have fully provided all the details of our proposed methods in the main text and appendix, to guarantee the reproducibility of our paper's main experimental results. Meanwhile, we will open-source all of our code, data, and models after our paper is accepted. These details and open-source initiatives will ensure the reproducibility of our work and make a significant contribution to the entire community.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: To address privacy concerns associated with our dataset containing persons and the application of person ReID technology, we will implement a controlled release of our code, data, and models. Therefore, to ensure private information security, we would not provide open access to our data and code during the paper submission process. We provide training and testing logs in the supplementary materials to ensure the authenticity and validity of our results. Notably, we will publicly share the application link for all our sources after our paper is accepted. Applicants will be required to adhere to relevant guidelines and regulations to access our sources, and we will conduct a strict review of their qualifications to prevent any privacy violations.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our paper specifies all the training and test details necessary to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

https://doi.org/10.52202/079017-0795

Answer: [No]

Justification: Due to pre-training technology is comparably computational expensive, the error bars are not reported in the paper like nearly all related work. We provide training and testing logs in the supplementary materials to ensure the authenticity and validity of our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our paper provides sufficient information on the computer resources needed to reproduce the experiments, which can be found in Sec. 4.1 and Sec. A.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper complies with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our paper discusses both potential positive societal impacts and negative societal impacts of the work performed in Sec. A.1 of the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our paper describes safeguards that have been put in place for responsible release of data and models in Sec. A.3 of the appendix.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

https://doi.org/10.52202/079017-0795

Answer: [Yes]

Justification: The creators and original owners of assets used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected, which can be found in Sec. A.4 of the appendix.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: To address privacy concerns associated with our assets including code, data, and models, we will implement a controlled release of our assets. Therefore, to ensure private information security, we would not release our assets at submission time. We will publicly share the application link for our assets after our paper is accepted. Applicants will be required to adhere to relevant guidelines and regulations to access our assets, and we will conduct a strict review of their qualifications to prevent any privacy violations.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-0795