Journal of Computer Science and Cybernetics, V.41, N.3 (2025), 265-284
DOI: 10.15625/1813-9663/23018

VITC-UREID: ENHANCING UNSUPERVISED PERSON REID WITH VISION TRANSFORMER IMAGE ENCODER AND CAMERA-AWARE PROXY LEARNING

DANG H. PHAM[1,3], TU N. NGUYEN[2], HOA N. NGUYEN[1,*]

[1] *VNU University of Engineering and Technology, 144 Xuan Thuy Street, Cau Giay Ward, Ha Noi, Viet Nam*
[2] *Kennesaw State University, 1000 Chastain Road, Kennesaw, GA 30144, USA*
[3] *University of Khanh Hoa, 01 Nguyen Chanh, Nha Trang Ward, Khanh Hoa, Viet Nam*

Abstract. Person re-identification (ReID) plays a crucial role in computer vision-based surveillance systems, enabling the accurate identification of individuals across multiple camera views. Traditional convolutional neural network (CNN)-based approaches, such as those utilizing ResNet-50, struggle to capture long-range dependencies and contextual relationships, limiting their effectiveness in diverse real-world scenarios. To overcome these challenges, recent advancements have explored Vision Transformer (ViT)-based architectures, leveraging self-attention mechanisms for enhanced feature representation. In this research, we introduce a ViT-based framework, namely ViTC-UReID, for unsupervised person ReID by incorporating a camera-aware proxy learning mechanism to improve feature consistency across different camera viewpoints. Moreover, ViTC-UReID also uses clustering algorithms to generate pseudo labels for samples in training datasets. Our approach significantly enhances cross-camera adaptation, reducing domain shift effects while maintaining strong feature discrimination. We evaluate our method on three widely used benchmarks Market-1501, MSMT17, and CUHK03, demonstrating its superior performance compared to existing state-of-the-art unsupervised methods, particularly those utilizing camera identity cues. Furthermore, our model achieves competitive accuracy with fully supervised methods, highlighting the effectiveness of transformer-based representations in complex person ReID scenarios. Our findings reinforce the growing potential of unsupervised person ReID methods and demonstrate that ViT architectures combined with camera-aware learning can drive substantial improvements in person ReID.

Keyword. Unsupervised person re-identification, enhanced image representation, camera-aware learning, vision transformer.

1. INTRODUCTION

Person re-identification (ReID) is a fundamental challenge in computer vision-based surveillance systems, involving the recognition of individuals across multiple security images or video frames captured from diverse viewpoints. As a critical component of intelligent

*Corresponding author.
E-mail addresses: dangph@vnu.edu.vn (D.H. Pham); tu.nguyen@kennesaw.edu (T.N. Nguyen); hoa.nguyen@vnu.edu.vn (H.N. Nguyen).

surveillance, person ReID supports applications such as threat detection, criminal identification, and multi-camera tracking. With the increasing deployment of surveillance infrastructure and heightened focus on public safety, the demand for robust person ReID solutions has surged, drawing significant attention from the research community.

ResNet-50 [1] has traditionally served as the backbone for image feature extraction in person ReID tasks, owing to its efficiency, stability, and strong benchmark performance. However, its limitations become evident in complex real-world scenarios: (i) It lacks the ability to model spatial relationships and long-range dependencies between body parts; (ii) It is not designed to capture semantic content or comprehend high-level contextual information; (iii) Its performance degrades under challenging conditions, such as low-light environments, occlusion, and variable camera angles. These limitations underscore the necessity for models with enhanced representational power and better generalization capabilities across diverse environments.

Initially, supervised person ReID methods employing deep network models were widely explored. However, models trained on publicly labeled datasets often struggle to perform effectively in specific real-world settings. Moreover, the escalating volume of data and the increasing time investment required for manual annotation pose substantial challenges in practical applications. To address these issues, researchers have shifted focus toward unsupervised person ReID approaches, which leverage unlabeled data for model training. These methods can be broadly categorized into two types: unsupervised domain adaptation (UDA) [2, 3] and fully unsupervised learning (USL) [4, 5]. UDA-based person ReID models employ an unlabeled source domain and a fully annotated target domain via transfer learning. However, reliance on the source domain can hinder model performance due to distribution discrepancies, adversely affecting knowledge transfer and effectiveness in the target domain. In contrast, USL-based person ReID methods exhibit greater flexibility and scalability, training directly on unlabeled datasets without external dependencies, making them more suitable for diverse real-world scenarios.

Modern USL person ReID frameworks integrate various components, including clustering algorithms [6, 7], memory banks [5, 8], and contrastive loss functions coupled with network models [4], leading to enhanced performance. Typically, USL person ReID involves generating pseudo-labels through clustering, computing contrastive loss using positive and negative memory bank samples, and iteratively updating cluster representation vectors to refine feature learning. However, a major challenge remains: effectively aligning images of the same individual despite pseudo-label noise.

Given these challenges, our contributions in this paper are threefold. First, we introduce a novel model based on the Vision Transformer architecture, which effectively captures both global contextual information and fine-grained local details. Unlike conventional convolutional networks, ViT employs self-attention mechanisms to extract meaningful representations, ensuring robust person ReID by leveraging both structural and contextual features. Second, we propose the integration of a camera-aware proxy learning mechanism within our model's training process. This approach mitigates domain shift issues stemming from variations in camera viewpoints, lighting conditions, and resolution inconsistencies. By incorporating camera-aware proxies, our method enhances feature consistency across different camera sources, thereby improving generalization and adaptability to real-world scenarios. Third, we conduct extensive evaluations on three widely recognized person ReID benchmarks Market-

1501, MSMT17, and CUHK03 to rigorously assess our approach. Comprehensive experiments demonstrate that our method significantly outperforms existing state-of-the-art techniques, reinforcing its effectiveness in tackling complex person ReID tasks.

The remainder of this paper is structured as follows: Section 2 presents the problem formulation and discusses prior work. Section 3 focuses on the proposed methodology in detail. Section 4 reports and analyzes the experimental findings. Finally, Section 5 concludes the paper and outlines directions for future work.

2. PROBLEM STATEMENT AND RELATED WORKS

2.1. Problem statement

In the USL person ReID problem, we train a model Ω on an unlabeled dataset containing N person images $\mathcal{D} = \{x_i\}_{i=1}^{N}$, aiming for high retrieval accuracy that can be deployed in real-world environments.

To evaluate the performance of person ReID models in real-world scenarios, we assess their accuracy on a test dataset $\mathcal{D}' = \{x'_i, y'_i\}_{i=1}^{M}$, where $\{x'_i, y'_i\}$ represents the i-th labeled sample, and M denotes the number of test samples. The USL person ReID problem can be formulated as follows

$$\textbf{score} = \mathcal{F}_{score}(\hat{\Omega}, \mathcal{D}') \text{ where } \hat{\Omega} = \mathcal{F}_{usl}(\Omega, \mathcal{D}). \tag{1}$$

Here, $\mathcal{F}_{score}$ represents the metric function, while $\mathcal{F}_{usl}$ denotes the unsupervised learning process. Assuming that a higher score corresponds to better model performance, our objective is to train models that achieve the highest possible score value.

2.2. Person ReID

In the early development of person ReID, hand-crafted features such as color histograms, texture descriptors, and local patterns (e.g., SIFT, LBP) were commonly paired with metric learning techniques. However, these methods struggled to cope with significant appearance variations caused by changes in pose, lighting, occlusion, and camera viewpoints. With the rise of deep learning, CNNs became the dominant approach due to their ability to automatically learn robust and discriminative visual features. CNNs effectively capture identity-related cues such as clothing, body shape, and color, enabling more accurate person matching. To further improve performance, researchers have explored various strategies. Specialized loss functions have been designed to enhance feature discriminability [9, 10], while others have proposed local feature learning techniques [11, 12] and attention mechanisms [13, 14, 15] to focus on salient regions of the person. Hybrid models combining both global and local representations have also shown strong performance [16, 17, 18]. In parallel, Generative Adversarial Networks (GANs) have been leveraged to synthesize cross-domain images and augment training data for person ReID [3, 19, 20]. Recently, transformer-based architectures and attention-driven models have gained popularity for their ability to model long-range dependencies. Notably, TransReID [21] improves robustness against appearance variations by leveraging global contextual information. Additionally, camera-aware learning methods such as CAP [22] address intra-class variance induced by camera discrepancies, significantly boosting unsupervised person ReID performance. Collectively, these advances have greatly improved the accuracy and robustness of modern person ReID models.

2.3. Vision transformers

Recently, Transformer-based architectures [23] have demonstrated superior performance in many computer vision tasks. Vision Transformer (ViT) [24] divides an image into patches, linearly projects them into tokens, and processes the sequence through a Transformer encoder with a learnable class token. Unlike CNNs, which rely on local receptive fields, ViTs leverage self-attention to capture long-range dependencies and global context. This capability is especially valuable for person ReID, where variations in viewpoint, illumination, pose, and occlusion often hinder recognition. Building on ViT, TransReID [21] adapts Transformer models for object ReID, introducing patch shuffle operations and incorporating side information such as camera and view IDs to enhance robustness. Other extensions integrate local or hybrid mechanisms: DCAL [25] employs a Transformer decoder to implicitly capture local features, while PAT [26] and HAT [27] combine CNNs and Transformers, with PAT using Transformers to generate attention masks and HAT aggregating hierarchical CNN features. Although these methods improve performance, most still lack explicit alignment for discriminative part-level features. To address this, PASS [28] proposes a pre-training framework with contrastive learning, introducing [PART] tokens to automatically extract local features. These studies highlight the growing potential of Transformer-based models in person ReID, where balancing global context and fine-grained part cues remains an open and promising research direction.

2.4. Image representations for person ReID

Global-level matching has long dominated person ReID, where image embeddings are compared in a shared space. Most methods employ ResNet backbones [1], applying global pooling to feature maps and optimizing with cross-entropy or triplet loss. While effective for overall appearance, these holistic features often miss fine-grained cues crucial for distinguishing similar identities. To address this, works such as CAMERA [29] adopt dilated convolutions, attention-based methods [30, 31] emphasize discriminative regions, and multi-granularity frameworks like MGN [32] and PPLR [16] combine global and local features. Meanwhile, ViT [24] excel at modeling global dependencies via self-attention, but may overlook local details since images are split into patch tokens. Recent works address this by enhancing local representation: TransReID [21] improves patch alignment, Token Labeling [33] supervises tokens directly, and multi-granularity or region-based frameworks [32] integrate global and local cues. More recently, NCL [34] proposes noisy-correspondence learning to strengthen patch-level alignments in ViT, proving that reliable local cues significantly boost discriminability. Building on these insights, our work leverages ViT with enhanced local correspondence modeling to balance holistic context and fine-grained details for robust person ReID.

3. PROPOSED METHOD

3.1. Approach direction

To tackle the problem described in previous sections, we introduce ViTC-UReID, a method that combines strong representation using a **ViT**-based backbone with **C**amera-aware Proxy Learning, enabling fully **U**nsupervised person **ReID** model training. The essence of ViTC-UReID lies in its integration of diverse representation techniques to generate more

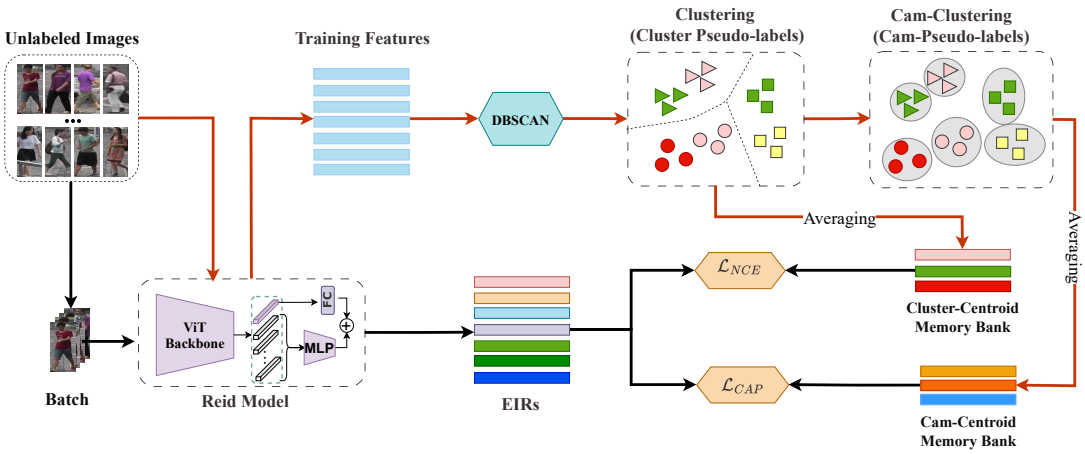


Figure 1: Overview of our proposed method, ViTC-UReID

discriminative features for retrieval. Furthermore, throughout training, the model learns to recognize discrepancies in images taken from different cameras, thereby reducing inter-camera variability. By incorporating camera-aware proxies, our approach improves feature alignment and enhances retrieval accuracy across different viewpoints.

Figure 1 visually depicts the core components of our method. Specifically, in the initial process (indicated by orange arrows), all images are passed through the ReID model to extract training features. These features are clustered using DBSCAN to generate initial cluster pseudo-labels. To further account for cross-camera variations, a camera-aware clustering step is applied, resulting in camera pseudo-labels. Based on these assignments, we compute cluster-centroids and cam-centroids by averaging the corresponding features. These centroids are then used to initialize the cluster-centroid memory bank and cam-centroid memory bank, respectively.

In the training process (indicated by **black arrows**), the pseudo-labeled data is divided into batches and passed through the ReID model to obtain updated embedding representations (EIRs). These EIRs are then compared with the centroids stored in the memory banks to compute two contrastive losses, $\mathcal{L}_{NCE}$ and $\mathcal{L}_{CAP}$. These losses jointly guide the model toward learning discriminative and camera-invariant representations.

The following subsections provide a comprehensive breakdown of each module within our framework.

3.2. Baseline model

Firstly, we leverage clustering algorithms to generate pseudo labels for training samples, enabling supervised model training without manual annotation. Specifically, given an unlabeled dataset $\mathcal{D}$, we apply DBSCAN [6] to assign pseudo labels to each sample x_i

$$\mathcal{C} = \text{DBSCAN}(\{\Omega(x_i) \mid x_i \in \mathcal{D}\}). \tag{2}$$

Here, $\Omega(x_i)$ represents the extracted image feature for sample x_i, while $\mathcal{C} = \{y_i\}_{i=0}^{N}$ denotes the pseudo labels assigned to the N samples in $\mathcal{D}$. DBSCAN groups the extracted features

into inliers (clustered samples) and outliers (unclustered samples). To ensure label reliability, outliers are discarded in subsequent training iterations, while clustered inliers are assigned to one of N centroids, forming pseudo labels that serve as supervision signals for loss computation and model optimization.

To train models for optimizing feature representations, the most recent unsupervised person ReID methods adopt contrastive learning, where the loss is computed within the context of a mini-batch. However, such mini-batch-based sampling inherently limits the diversity of positive and negative pairs, leading to suboptimal training signals. To overcome the limitation of locally selecting positive and negative samples, SpCL [4] stores feature representations in a global memory and updates them progressively during training. However, batch training only updates a subset of instances per iteration, causing an imbalanced updating pace and shifting the feature distribution. Cluster Contrast [35] addresses this issue by first calculating the cluster centroids by the mean feature vector of each cluster as

$$c_k = \frac{1}{|C_k|} \sum_{f_i \in C_k} f_i, \tag{3}$$

where f_i is the feature vector of example i, and C_k denotes the k-th cluster set and $|\cdot|$ indicates the number of feature vector i per cluster.

It then utilizes a **memory bank**, which refers to a structure designed to store and manage encoded feature representations in a flexible and efficient manner. Serving as a dynamic dictionary, the memory bank provides a large and diverse set of features for contrastive learning. By maintaining this structure, the model can compute contrastive loss more effectively without relying on large batch sizes, while also improving training stability and feature discrimination.

In our framework, the centroids of the feature clusters are stored in a cluster-centroid memory bank, which extends the memory bank concept to the cluster level. These centroids act as stable supervisory targets and support a momentum update mechanism, guiding the gradual refinement of feature embeddings across training epochs. By leveraging memory at the cluster level, the model benefits from more consistent updates and improved robustness throughout the USL process.

During the learning process, the ClusterNCE loss [35] is employed to compute the similarity between the query feature q and all cluster centroids. The formula for the loss function is given by

$$\mathcal{L}_{NCE} = -\log \frac{\exp(q \cdot c_+/\tau)}{\sum_{k=1}^{K} \exp(q \cdot c_k/\tau)}, \tag{4}$$

where c_+ indicates the positive centroid class, or the cluster which q is belong to, c_k is the representation vector of the k-th cluster, τ is the temperature hyper-parameter.

This loss function guides the model to align each query feature with its corresponding cluster centroid, thereby improving the consistency and discriminability of the learned representations. To ensure stable and continuous refinement of cluster centroids, we update each centroid using the query feature through a momentum-based update

$$c_k \leftarrow m \cdot c_k + (1 - m) \cdot q, \tag{5}$$

with c_k is cluster centroid of cluster k, q is the query feature and m is momentum update. The parameter m regulates the consistency between the cluster feature and the most recently

observed query instance. As m approaches 0, the updated cluster centroid becomes more aligned with the newest query features. This approach enables effective representation learning without the need for annotated labels, making it well-suited for practical applications where labeled data is limited.

3.3. Enhanced image representation

Initially, the baseline architecture to be discussed is depicted in Figure 1, which follows the conventional ViT backbone network to derive representations for images [24]. As mentioned in Section 2, the image encoder is initiated using the pre-trained model LUPerson-ViT [36]. In detail, given an input image I that belongs to $R^{(H\times W\times C)}$, it is divided into non-overlapping fragments $M = H \times W/S^2$, where S refers to the size of each fragment. These fragments are subsequently linearly embedded and hence, are learnable. A [CLS] token is appended at the start to represent the image-level. Following this, the patch set $P = \{p_{cls}, p_1, p_2, ..., p_{M-1}, p_M\}$ is introduced to the P transformer blocks of the image encoder. This action yields a sequence of D-dimensional representations of the image

$$F' = \{f'_{cls}, f'_1, f'_2, ..., f'_{M-1}, f'_M\} = \text{Transformer Block (P)}. \tag{6}$$

These representations are then fed into a full-connection layer to obtain the fine-grained features as

$$F = \{f_{cls}, f_1, f_2, ..., f_{M-1}, f_M\}, \text{where } f_i = FC(f'_i) \text{ with } f'_i \in F'. \tag{7}$$

The global visual representation of the entire image is considered to be f^I_{cls}, while the local features of the patch are denoted as $\{f_j|(i = 1, 2, ..., M)\}$. In both training and inference, only the global features are utilized to calculate loss values and compute the distances between images for retrieval. Although global features provide a high-level summary of an image's content and context, including primary objects and scene layout, they could ignore fine-grained details from local features, such as textures and colors, that play a crucial role in precise object recognition [34].

Motivated by that, we propose using fused features, which also aggregate crucial local features from informative tokens and add this information to the global feature. For instance, as shown in Section 1, given an image I, we can obtain two feature sets F and F' as well as the self-attention map $\mathbf{A} \in \mathbb{R}^{(1+M)\times(1+M)}$ from the last layer of the encoder. The correlation scores between the global token [CLS] and local tokens are denoted as $m = \mathbf{A}[0, 1:] \in \mathbb{R}^M$. We then select the top $\mathbf{K}$ percent of features with higher scores from the M local embeddings in feature set F' as

$$\mathbf{F}^* = \{f'_j \text{ if } m_j \text{ in top-K}\}. \tag{8}$$

Finally, we enhance the feature representation with a multi-layer perceptron (MLP) layer (including only two fully connected layers) as

$$V = \text{MaxPool}(MLP(\|F^*\|) \oplus f_{cls}. \tag{9}$$

where $||*||$ denotes the operator norm and $\oplus$ is the concatenation operator. We employ max pooling to refine local information and scale it to vectors with the same dimension as the set F. According to that, we replace the original global features f_{cls} in loss computation and retrieval by the enhanced features V.

3.4. Camera-aware proxy learning

The core challenge in person ReID is matching people across different cameras, so inspired by Wang et al. [22], we focus on inter-camera learning, which tries to match the same person across different cameras. Specifically, we apply camera-aware proxy, which is defined by splitting each identity cluster into multiple subgroups based on CamID. Each subgroup represents the same identity from a single camera. This means building more reliable pseudo labels, since samples from the same camera tend to have less appearance variance.

Specifically, for each dataset, we have a set of cameras denoted by $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$. We denote a sample x that belongs to both cluster a and camera b as x_a^b. The representation embedding vector of x_a^b is denoted as f_a^b. We define a *camera-aware proxy* c_a^b as the centroid of features corresponding to all samples from cluster a captured by camera b

$$c_a^b = \frac{1}{|X_a^b|} \sum_{x \in X_a^b} f(x), \tag{10}$$

where X_a^b denotes the set of all samples in the cluster a captured by the camera b, and $f(x)$ is the feature embedding of the sample x. Like cluster centroids, these camera-aware proxies are also stored in a memory structure referred to as the cam-centroid memory bank. From that, the inter-camera contrastive loss is a softmax log loss of query q with one positive cross-camera proxy and N_{neg} hard negatives in the memory proxies, as follows

$$\mathcal{L}_{CAP} = -\frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp(q \cdot c_a^j / \tau_c)}{\exp(q \cdot c_a^j / \tau_c) + \sum_{k \in \mathcal{N}(i)} \exp(q \cdot c_k / \tau_c)}, \tag{11}$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the index sets of the positive and hard negative proxies, respectively, τ_c is the temperature hyper-parameter.

3.5. Overall objective function

In summary, we train our model that generates enhanced image representations by Equation 9. Based on these features, the accuracy and reliability of pseudo labels could be improved, and the retrieval performance is also enhanced. For model optimization, we use ClusterNCE and CAP loss functions through training (Equation 4 and Equation 11, respectively). The overall objective of training can be written as follows

$$\mathcal{L}_{OBJ} = \mathcal{L}_{NCE} + \lambda * \mathcal{L}_{CAP}. \tag{12}$$

Here λ is a coefficient used to control the impact of CAP in training. Besides, we also leverage the benefits of data augmentation to enhance the robustness and accuracy of person ReID models. The entire training process is shown in Algorithm 1.

4. EXPERIMENTS AND EVALUATION

To demonstrate the performance of our ViTC-UReID method, we conduct a comprehensive experiment to analyze the effectiveness of our proposals. Following these objectives, we first present the experimental settings. Secondly, we discuss several facets of our proposed method across multiple benchmarks. Finally, we compare the performance of our approach against several state-of-the-art unsupervised person ReID methods.

Algorithm 1: ViTC-UReID: Enhancing unsupervised person ReID with ViT image encoder and camera-aware proxy learning

Input: Training dataset $D = \{I_i\}_{i=1}^{M}$; a model Θ; hyper-parameters: λ, *top-K*
Output: Trained person ReID model Θ with optimized parameters

Initialize Θ with the weights of the pre-trained LUPerson;
while *not* Θ *converged* **do**
  Extract features: $V \leftarrow \Theta(D)$;
  Generate cluster pseudo labels: $L \leftarrow \text{DBSCAN}(V)$;
  Generate camera pseudo labels: $P \leftarrow \text{CamCluster}(L)$;
  Initialize cluster-centroid memory bank: $M \leftarrow \text{Init}(V, L)$;
  Initialize cam-centroid memory bank: $C \leftarrow \text{Init}(V, P)$;
  for *each* $x = \{(I_i, c_i, L_i, P_i)\}_{i=1}^{B}$ *in* $zip(D, L, P)$ **do**
    $V_i \leftarrow \Theta(I_i)$;
    $\mathcal{L}_1 \leftarrow \mathcal{L}_{\text{NCE}}(V_i, L_i, M)$;
    $\mathcal{L}_2 \leftarrow \mathcal{L}_{\text{CAP}}(V_i, c_i, C)$;
    $\mathcal{L}_{\text{OBJ}} \leftarrow \mathcal{L}_1 + \lambda \cdot \mathcal{L}_2$;
    $\Theta \leftarrow \text{Optimizer}(\Theta, \mathcal{L}_{\text{OBJ}})$;
  end
  Evaluate the model;
end

4.1. Datasets

We use three popular datasets Market-1501 [37], MSMT17 [38], and CUHK03 [39] to evaluate our method.

- Market-1501: Includes 32,668 images of 1,501 identities from six campus cameras. Images are pre-cropped using the DPM detector. It's widely used due to its scale and real-world relevance.
- MSMT17: A large and challenging dataset with 126,441 images of 4,101 identities from 15 cameras (12 outdoor, 3 indoor). It offers high diversity in environment, weather, and time, ideal for testing model generalization.
- CUHK03: Contains 13,164 images of 1,467 identities from six cameras at the Chinese University of Hong Kong. It includes both manually cropped (labeled) and detector-cropped (detected) versions to assess model sensitivity to detection quality. In our work, we evaluate on the labeled set.

We refer to these datasets "Market-1501" as "Market" as, "MSMT17" as "MSMT" and "CUHK03" as "CUHK" throughout the paper.

4.2. Evaluation metrics

To evaluate the efficacy of the proposed method, we utilize mean average precision (mAP) and Cumulative Matching Characteristic at Rank@n (CMC $R@n$) metrics. Mean Average Precision (mAP) is frequently employed to evaluate a model's proficiency in detecting and

classifying objects, especially in object detection tasks. The calculation involves averaging the average precision (AP) values across all labels, with each AP obtained from the areas under the precision-recall curve. The mAP formula is presented in Equation 13

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{k=1}^{M_i}(P(k) * \text{rel}(k)), \tag{13}$$

where N denotes the total number of queries, M_i signifies the total number of predictions for a particular label; $P(k)$ represents the precision at cut-off k, defined as the ratio of true positive items within the top k retrieved items; $\text{rel}(k)$ is an indicator function that equals 1 if the item at position k is a true positive and 0 otherwise.

CMC Rank@n, specifically R@1/5/10, serves as a prevalent evaluation metric for retrieval tasks. This metric assesses the probability of identifying an accurate match among the top-n ranked outcomes. Within the realm of retrieval, CMC Rank@n (R@n) is essential as it assesses the efficacy of a model in recognizing an individual from a gallery of images using a query image. The formula for calculating the R@n metric is presented in Equation 14

$$R@n = \frac{\text{Number of correct matches within top-n}}{\text{Total number of query images}} \times 100\%. \tag{14}$$

The evaluation protocol focuses on retrieving images that match textual descriptions of individuals, primarily aiming to maximize the Rank@1 score, while mAP functions as a supplementary performance metric.

4.3. Experiment setup

Our method standardizes the input image to a resolution of 256×128 pixels. For data augmentation, we employ techniques such as Random Horizontal Flipping, Random Cropping with Padding, and Random Erasing. The model architecture is based on ViT-B/16 as the image encoder and initialized with pretrained LUPerson [36]. Throughout the training process, we utilize the SGD optimizer for 60 epochs with a batch size of 128 and an initial learning rate of 1×10^{-3}. The learning rate undergoes a tenfold reduction at the milestone epochs 30 and 50.

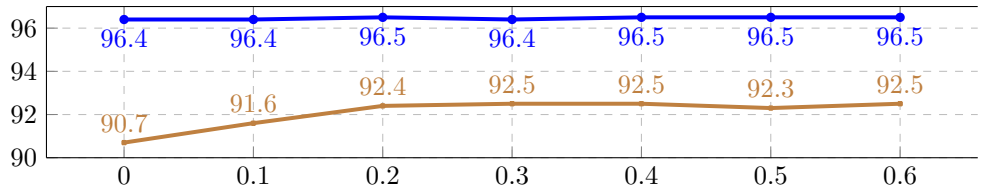
All of our experiments are conducted on a machine with a single NVIDIA RTX 4060 Ti with 16GB memory, Intel Core™ i5-13600K, and 64GB RAM. Besides that, we implement our method using Pytorch v1.12.1 with Python 3.8.18.

4.4. Ablation studies

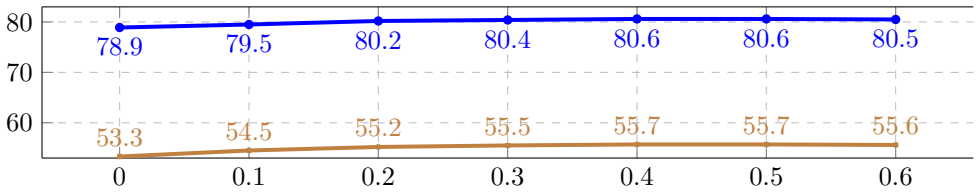
Analysis of Baseline: Initializing the ViT backbone with pre-trained weights from the large-scale unlabeled dataset LUPerson [36] has proven to be an effective approach in person ReID tasks. The pre-trained ViT backbone acts as a powerful feature extractor, capable of capturing fine-grained visual details from human images, which significantly enhances the model's robustness and generalization ability. Similar strategies have been employed in works like TransReID-SSL [31] and TMGF [40], demonstrating that leveraging pre-trained backbones is a widely accepted technique. However, to assess the impact of this initialization, we establish a baseline by evaluating the model's performance when initialized

Table 1: Ablation studies on Market and MSMT

Method	Market				MSMT			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Pretrained	34.2	49.7	57.3	11.3	20.5	29.5	34.1	4.7
Baseline	96.4	98.8	99.4	90.7	78.9	88.0	90.8	53.3
Baseline + *EIR*	96.5	98.8	99.5	92.5	80.6	88.7	91.3	55.7
Baseline + *CAP*	96.6	98.9	99.5	92.0	85.4	92.1	94.0	62.9
Baseline + *EIR* + *CAP*	97.1	99.1	99.5	92.8	85.8	92.3	94.1	63.6



(a) Market



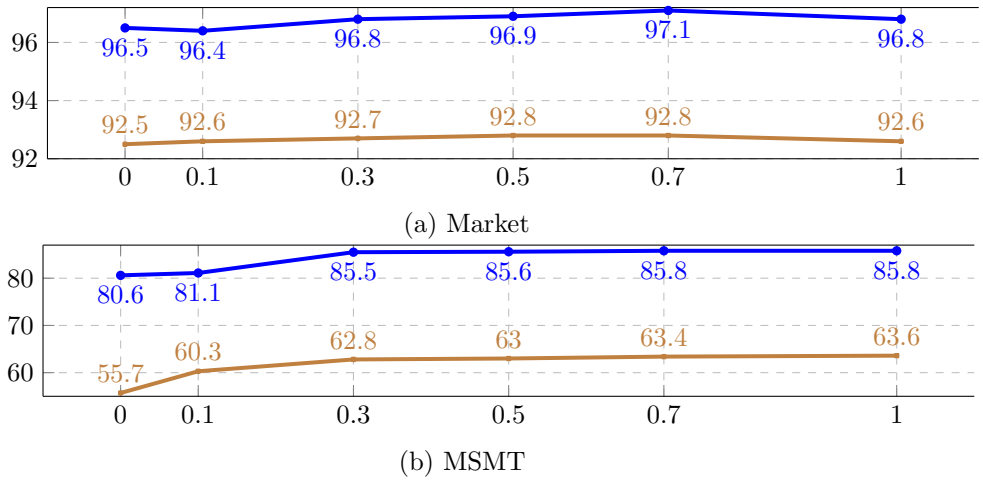
(b) MSMT

Figure 2: Performance R@1 and *mAP* under different top-K values on Market and MSMT

with LUPerson's pre-trained weights before fine-tuning. The initial results, as seen in Table 1, show that while the *Pretrained* provides a strong starting point, it performs poorly when directly applied to domain-specific datasets, achieving only 34.2% R@1 and 11.3% mAP on Market, and 20.5% R@1 and 4.7% mAP on MSMT. This underscores the necessity of fine-tuning in adapting the model to specialized domains.

Next, we fine-tune the initialized model using the basic approach using DBSCAN and train with only $\mathcal{L}_{NCE}$ as described in Section 3.2. The results, considered as the **baseline**, are presented in the second row of Table 1, highlighting the crucial role of domain adaptation. Following fine-tuning, the model achieves significant performance improvements, reaching 96.4% R@1 and 90.7% mAP on Market, and 78.9% R@1 and 53.3% mAP on MSMT. These results confirm that while pre-training provides the model with generalized feature extraction capabilities, fine-tuning is essential for optimizing performance within specific domains. Through this adaptation process, the model refines its feature representations, aligning them with dataset-specific characteristics to enhance retrieval accuracy. Thus, the findings reinforce that fine-tuning plays a vital role in developing a high-performing model tailored to a particular domain. Relying solely on knowledge transferred from a large-scale pre-trained backbone proves insufficient for achieving optimal performance, underscoring the necessity of domain adaptation.

Different combinations of the components: In this section, we assess the effectiveness



(a) Market

(b) MSMT

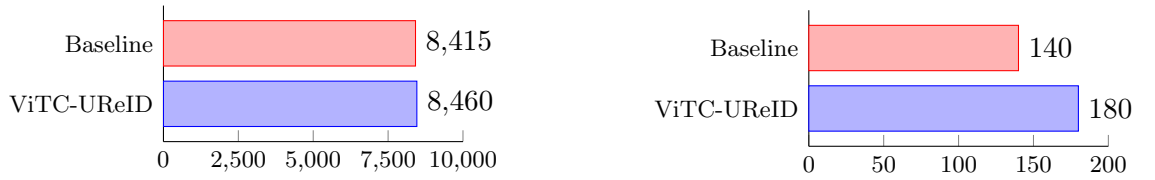
Figure 3: Performance R@1 and mAP under different λ on Market and MSMT

of the proposed components, beginning with the EIR module. As demonstrated in Table 1, integrating EIR enhances model performance by up to 2% in R@1 and mAP. Notably, the most significant improvement is observed on the MSMT benchmark. Unlike the Market dataset, MSMT consists of images captured under diverse environmental conditions, leading conventional models to overlook essential features. This variability can hinder accurate identity retrieval, as models may fail to focus on the most relevant attributes of an individual's appearance. By incorporating EIR, the most crucial features of a person's image are reinforced within the feature representation, ensuring a balanced integration of both local and global information. Consequently, this enhancement significantly improves retrieval precision across challenging scenarios.

Besides, we also present the best *top-K percent of local features* used to generate EIR in Figure 2. The figure shows that EIR stabilizes when top-K exceeds 0.2. If we set a small top-K (top-K < 0.2), it could lack sufficient information to make an improvement. Otherwise, using a high top-K (≥ 0.5) adds unnecessary complexity in calculation, while the performance is not improved significantly. Based on findings, we set top-K to 0.4 to achieve optimal performance with reasonable training time.

Similarly, we fine-tune models without EIR but with the addition of CAP, and we present the results in the fourth row of Table 1. Like EIR, CAP contributes a slight performance increase on the Market dataset; however, it leads to a significant improvement on the MSMT benchmark. This observation underscores the importance of CamID as a crucial factor in enabling the model to learn more discriminative representations. By incorporating CAP, the model is better equipped to account for variations in camera viewpoints, enhancing its ability to differentiate individuals effectively.

Finally, we integrate both EIR and CAP into the training process. Compared to the original baseline, the final model demonstrates a noticeable performance boost. Specifically, we achieve 97.1% and 92.8% in R@1 and mAP on the Market benchmark, while on MSMT, the model attains 85.8% and 63.6% in R@1 and mAP. To achieve these outcomes, we have to adjust the value λ to control the effect from CAP. As illustrated in Figure 3, the best values are 0.7 and 1.0, leading to the most optimal and stable performance on Market and MSMT datasets, respectively. On the other hand, these results also validate the effectiveness of



(a) GPU Memory Usage (MB)

(b) Training Time per Epoch (s)

Figure 4: Computational efficiency: Baseline vs ViTC-UReID on Market

combining EIR and CAP, highlighting their complementary roles in improving representation learning and retrieval accuracy across diverse datasets.

4.5. Computational efficiency

To further assess the practical impact of our method, we evaluate the computational efficiency of ViTC-UReID compared to the baseline approach on the Market dataset using a batch size of 128 as a representative case. While the introduction of EIR and CAP significantly improves feature discriminability and overall person ReID performance, it also incurs additional computational overhead during training. Using PyTorch's profiling tools, we measure peak GPU memory consumption and training time per epoch.

As shown in Figure 4, GPU memory usage remains nearly identical between the two methods, with ViTC-UReID consuming 8,460 MB, only slightly higher than the baseline's 8,415 MB. In terms of training time, ViTC-UReID requires 180 seconds per epoch compared to 140 seconds for the baseline, representing a 28.5% increase. This overhead primarily stems from the integration of local feature enhancement within the ViT encoder and the camera-aware proxy learning mechanism.

Despite the additional training cost, the performance gains in mAP and R@n accuracy clearly justify the trade-off. Furthermore, the enhanced training stability and improved representation learning achieved by ViTC-UReID are crucial for robust and generalizable person ReID, especially in cross-camera and real-world deployment scenarios.

4.6. Visualization analysis

To provide a more intuitive demonstration of ViTC-UReID's effectiveness, we conduct multiple visualization analyses. In Figure 5, we utilize t-SNE to visualize the learned features of the six identities with the highest sample count in the MSMT dataset. Each point is color-coded, with identical colors indicating samples belonging to the same identity. As shown in Figure 5(b), our model effectively clusters features corresponding to the same identity while maintaining clear separation between images of different identities. In contrast, Figure 5(a) illustrates that the baseline model struggles with identity separation, often grouping features from different identities together, making distinguishing between individuals more challenging. On the other hand, these figures also indicate noise in pseudo labels, which is an inevitable challenge in clustering-based USL person ReID.

We present representative retrieval results to qualitatively assess the effectiveness of our proposed method. Figure 6 showcases the top 10 retrieval results, highlighting both successful and failure cases in comparison to the baseline model. On the left, the results

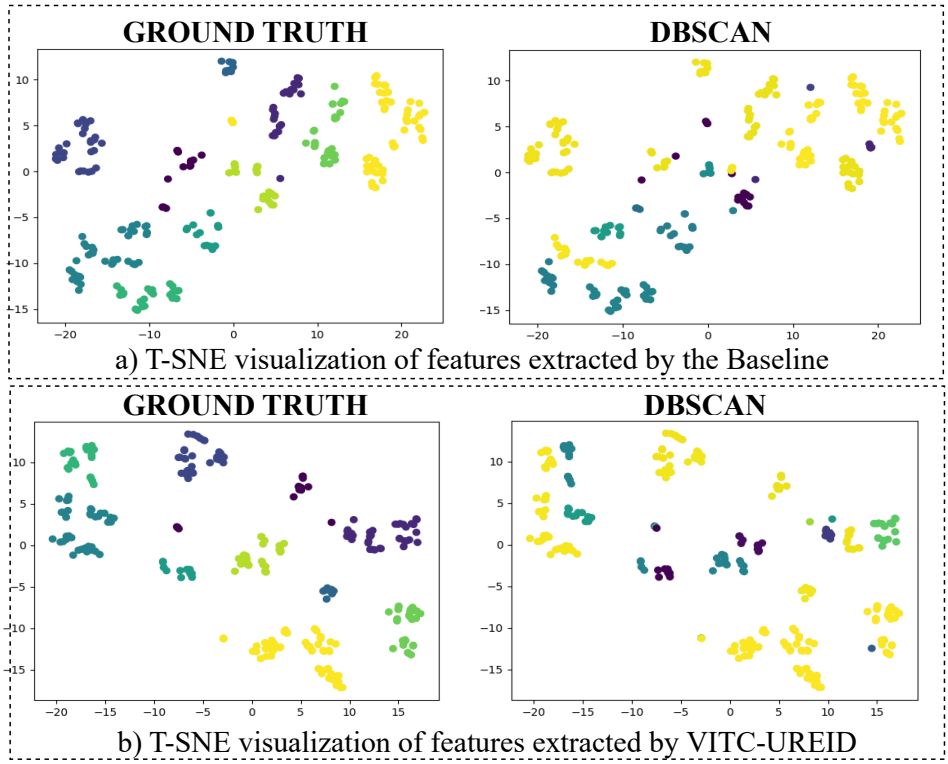


Figure 5: T-SNE visualization of features on MSMT

Figure 6: A comparison of the ten highest retrieval results on the MSMT dataset. We use the baseline model (a) and the ViTC-UReID model (b) for each query.

from the baseline model are displayed, while the right side presents the retrievals generated by ViTC-UReID, with correctly retrieved images marked by green boxes. In the majority of cases, our model successfully retrieves images of the targeted individual, demonstrating superior ranking capability. Notably, ViTC-UReID excels in ranking the correct matches higher than other candidates, reinforcing its effectiveness in person ReID. However, it is worth mentioning that some retrieved images, despite belonging to different identities, bear a strong resemblance to the query and appear visually plausible.

Table 2: Comparison with SOTA methods. [Keys: **The best of unsupervised methods**, † indicates methods using the camera information].

Back bone	Method	Venue	Market				MSMT				CUHK			
			R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Fully supervised methods														
CNN	ABNET+	CVPR22	95.7	-	-	93.0	80.8	-	-	62.2	80.6	-	-	79.1
	NFormer [41]													
	ProNet++ [42]	ArXiv23	96.0	–	–	90.2	85.4	–	–	65.5	85.2	–	–	82.7
ViT	TransReID [21]	ICCV21	95.2	–	–	89.5	86.2	–	–	69.4	–	–	–	–
	CLIP-ReID	AAAI23	95.5	–	–	89.6	88.7	–	–	73.4	–	–	–	–
	TMGF† [40]	WACV23	96.3	98.9	99.3	91.9	88.2	94.1	95.4	70.3	–	–	–	–
Fully unsupervised methods														
CNN	CC [35]	ACCV22	92.9	97.2	98.0	83.0	62.0	71.8	76.7	33.0	–	–	–	–
	PPLR† [16]	CVPR22	94.3	97.8	98.6	84.4	73.3	83.5	86.5	42.2	–	–	–	–
	ISE [43]	CVPR22	94.3	98.0	98.8	85.3	67.6	77.5	81.0	37.0	–	–	–	–
ViT	PASS [28]	ECCV22	94.9	-	-	88.5	67.0	-	-	41.0	–	–	–	–
	PCL-CLIP† [44]	Arxiv23	94.8	98.0	98.7	88.4	84.9	92.0	94.0	**65.5**	–	–	–	–
	TMGF† [40]	WACV23	95.5	98.0	98.7	89.5	83.3	90.2	92.1	58.2	–	–	–	–
	ACFL-ViT [45]	PR24	95.1	-	-	89.1	70.1	-	-	45.7	–	–	–	–
	TCMM [46]	Arxiv25	96.0	-	-	90.5	78.4	-	-	52.0	–	–	–	–
ViT	ViTC-UReID	(our)	**97.1**	**99.1**	**99.3**	**92.8**	**85.8**	**92.3**	**94.1**	63.6	**91.1**	**95.1**	**97.1**	**89.8**

4.7. Comparison with state-of-the-art methods

Our final experimental results, as displayed in Table 2, demonstrate that our method achieves superior performance across multiple evaluation metrics when compared to several well-known methods.

Comparison between CNN-based and ViT-based unsupervised methods. On Market, the unsupervised methods based on CNN architectures (eg., CC, PPLR, and ISE) generally yield R@1 accuracies in the low-to-mid 90s with mAP values ranging from 83.0% to 85.3%. In contrast, the unsupervised methods built on ViT such as TransReID [21], PASS [28], PCL-CLIP [44], TMGF [40], ACFL-ViT [45], TCMM [46] tend to achieve slightly higher R@1 accuracies (around 94.8% to 96.0%) and notably improved mAP scores (from 88.4% up to 90.5% on Market). On MSMT, this trend is even more pronounced: while CNN-based methods report rather modest mAP values (e.g., PPLR at 42.2%), several ViT-based approaches push the performance significantly higher. These results indicate that ViT-based representations can better capture global and contextual information, thus offering a clear performance edge in unsupervised person ReID tasks.

Comparison of unsupervised methods incorporating CamID information: A number of unsupervised approaches incorporate CamID cues to handle cross-view inconsistencies, as denoted by the dagger (†) symbol. In the CNN-based category, PPLR demonstrates competitive performance with a R@1 accuracy of 94.3% and an mAP of 84.4% on Market. On the ViT side, two methods TMGF and PCL-CLIP exploit CamID information, with TMGF achieving an R@1 of 95.5% and mAP of 89.5%, and PCL-CLIP further enhancing performance on challenging datasets. Compared to them, our ViTC-UReID also utilizes CamID information, and reaches the highest performance with 97.1% and 92.8% in R@1 and mAP on Market, as well as obtains an R@1 of 85.8% and mAP of 63.6%. These observations confirm that

the integration of CamID information within unsupervised frameworks improves feature discrimination across different camera views, which is especially beneficial when handling diverse data distributions such as those in MSMT.

Comparison with fully supervised methods: When compared to their unsupervised counterparts, fully supervised methods continue to hold an advantage across several benchmarks. In the CNN-based group, methods like ABNET+NFormer [41] achieve a R@1 accuracy of 95.7% with an impressive mAP of 93.0% on Market. Similarly, among the ViT-based supervised approaches, TMGF (with supervision) and TransReID maintain high performance, with TMGF reporting a R@1 of 96.3% and mAP of 91.9% on Market. Although these supervised methods tend to deliver superior results due to the availability of annotated data, the performance gap between our model and them is narrowing as we leverage ViT architectures and additional CamID cues, and continue to evolve. Apparently, with the ViT-backbone-based model, our method greatly outperforms CNN-based methods, even when training in an unsupervised manner.

5. CONCLUSION

In this work, we have addressed the enduring challenges of person ReID, which include handling diverse real-world scenarios, coping with complex variations in camera viewpoints, and overcoming the limitations of conventional backbone networks such as ResNet-50. The key challenges lie in effectively capturing long-range spatial dependencies, modeling high-level semantic context, and mitigating the inevitable domain shifts arising from varying conditions such as illumination, occlusion, and low resolution.

Our proposal introduces an unsupervised framework built upon the ViT architecture, which significantly enhances feature representation by harnessing self-attention mechanisms to capture both global contextual and fine-grained local details. We further integrate a camera-aware proxy learning strategy to alleviate cross-view discrepancies, thereby promoting feature consistency across heterogeneous camera sources. Extensive evaluations on benchmarks such as Market-1501, MSMT17, and CUHK03 reveal that the proposed approach not only surpasses several state-of-the-art unsupervised methods, especially those incorporating CamID cues, but also starts to close the performance gap with fully supervised techniques.

While our method demonstrates clear advantages in robustness and adaptability, several limitations remain. The reliance on high-quality pseudo-label generation may still incur noise, particularly in highly cluttered or low-quality visual environments. Moreover, the computational demands of transformer architectures limit deployment in resource-constrained settings. Future work will focus on improving pseudo-label reliability, exploring lightweight architectures for real-time applications, and extending the camera-aware proxy learning mechanism to better handle extreme environmental variations. These directions promise to further bridge the gap between unsupervised and supervised person ReID performance, paving the way for more versatile and scalable surveillance solutions.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[2] Y. Tu, "Domain camera adaptation and collaborative multiple feature clustering for unsupervised person re-id," in *Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis*. ACM, 2022, pp. 51–59.

[3] D. H. Pham, A. D. Nguyen, and H. N. Nguyen, "Gan-based data augmentation and pseudo-label refinement with holistic features for unsupervised domain adaptation person re-identification," *Knowledge-Based Systems*, vol. 288, p. 111471, 2024.

[4] Y. Ge, F. Zhu, D. Chen, R. Zhao, and h. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 11 309–11 321.

[5] Z. Hu, C. Zhu, and G. He, "Hard-sample guided hybrid contrast learning for unsupervised person re-identification," in *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, 2021, pp. 91–95.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, pp. 226–231.

[7] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 654–13 662.

[8] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.

[9] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. Washington D.C, USA: AAAI Press, 2017, pp. 2194–2200.

[10] M. Wieczorek, B. Rychalska, and J. Dabrowski, "On the unreasonable effectiveness of centroids in image retrieval," in *Neural Information Processing*, T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, Eds. Cham: Springer International Publishing, 2021, pp. 212–223.

[11] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2197–2206.

[12] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. New York City, USA: IEEE, 2019, pp. 542–551.

[13] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New York City, USA: IEEE, 2018, pp. 2119–2128.

[14] G. Zhang, Y. Zhang, T. Zhang, B. Li, and S. Pu, "Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 133–14 142.

[15] B. Yang, Y. Shan, R. Peng, J. Li, S. Chen, and L. Li, "A feature extraction method for person re-identification based on a two-branch CNN," *Multimedia Tools and Applications*, vol. 81, no. 27, pp. 39 169–39 184, Nov. 2022.

[16] Y. Cho, W. J. Kim, S. Hong, and S.-E. Yoon, "Part-based pseudo label refinement for unsupervised person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7298–7308.

[17] J. Xi, J. Huang, S. Zheng, Q. Zhou, B. Schiele, X.-S. Hua, and Q. Sun, "Learning comprehensive global features in person re-identification: Ensuring discriminativeness of more local regions," *Pattern Recognition*, vol. 134, p. 109068, 2023.

[18] D. Zhang, H. Fan, X. Zhou, and L. Su, "Joint global feature and part-based pyramid features for unsupervised person re-identification," *Journal of Electronic Imaging*, vol. 33, no. 2, p. 023043, 2024.

[19] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[20] T. Wu, R. Zhu, and S. Wan, "Semantic map guided identity transfer gan for person re-identification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 11, Sep. 2024.

[21] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14 993–15 002.

[22] M. Wang, B. Lai, J. Huang, X. Gong, and X.-S. Hua, "Camera-aware proxies for unsupervised person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, pp. 2764–2772, May 2021.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[25] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4682–4692.

[26] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2897–2906.

[27] G. Zhang, P. Zhang, J. Qi, and H. Lu, "Hat: Hierarchical aggregation transformers for person re-identification," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 516–525.

[28] K. Zhu, H. Guo, T. Yan, Y. Zhu, J. Wang, and M. Tang, "Pass: Part-aware self-supervised pre-training for person re-identification," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*. Berlin, Heidelberg: Springer-Verlag, 2022, pp. 198–214.

[29] L. Qu, M. Liu, D. Cao, L. Nie, and Q. Tian, "Context-aware multi-view summarization network for image-text matching," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1047–1055.

[30] J. Ding and X. Zhou, "Learning feature fusion for unsupervised domain adaptive person re-identification," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 2613–2619.

[31] H. Luo, P. Wang, Y. Xu, F. Ding, Y. Zhou, F. Wang, H. Li, and R. Jin, "Self-supervised pre-training for transformer-based person re-identification," *arXiv preprint arXiv:2111.12084*, 2021.

[32] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 274–282.

[33] Z.-H. Jiang, Q. Hou, L. Yuan, D. Zhou, Y. Shi, X. Jin, A. Wang, and J. Feng, "All tokens matter: Token labeling for training better vision transformers," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 18 590–18 602.

[34] Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, and P. Hu, "Noisy-correspondence learning for text-to-image person re-identification," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27 187–27 196.

[35] Z. Dai, G. Wang, W. Yuan, S. Zhu, and P. Tan, "Cluster contrast for unsupervised person re-identification," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, December 2022, pp. 1142–1160.

[36] B. Hu, X. Wang, and W. Liu, "Personvit: large-scale self-supervised vision transformer for person re-identification," *Mach. Vision Appl.*, vol. 36, no. 2, Jan. 2025.

[37] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.

[38] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88.

[39] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. New York City, USA: IEEE, jun 2014, pp. 152–159.

[40] J. Li, M. Wang, and X. Gong, "Transformer based multi-grained features for unsupervised person re-identification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, January 2023, pp. 42–50.

[41] H. Wang, J. Shen, Y. Liu, Y. Gao, and E. Gavves, "Nformer: Robust person re-identification with neighbor transformer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7287–7297.

[42] Q. Wang, X. Qian, B. Li, Y. Fu, and X. Xue, "Rethinking person re-identification from a projection-on-prototypes perspective," 2023.

[43] X. Zhang, D. Li, Z. Wang, J. Wang, E. Ding, J. Shi, Z. Zhang, and J. Wang, "Implicit sample extension for unsupervised person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2022, pp. 7359–7368.

[44] J. Li and X. Gong, "Prototypical contrastive learning-based clip fine-tuning for object re-identification," 2023. [Online]. Available: https://arxiv.org/abs/2310.17218

[45] H. Ji, L. Wang, S. Zhou, W. Tang, N. Zheng, and G. Hua, "Transfer easy to hard: Adversarial contrastive feature learning for unsupervised person re-identification," *Pattern Recognition*, vol. 145, p. 109973, 2024.

[46] Z.-A. Zhu, H.-C. Chien, and C.-K. Chiang, "Tcmm: Token constraint and multi-scale memory bank of contrastive learning for unsupervised person re-identification," 2025. [Online]. Available: https://arxiv.org/abs/2501.09044

Received on June 11, 2025
Accepted on July 04, 2025