Revisiting Person Re-ID: ConvNeXt with AIBN and TNorm in IICS/IIDS Frameworks

Faisal Z. Qureshi[a] and Roya Dehghani
Faculty of Science, Ontario Tech University, 2000 Simcoe St. N., Oshawa, Ontario L1G 0C5, Canada
faisal.qureshi@ontariotechu.ca, roya.dehghani@ontariotechu.net

Keywords: Person ReID, ConvNeXT, Surveillance

Abstract: This paper investigates the integration of ConvNeXt—a convolutional architecture inspired by vision transformers—into the Intra- and Inter-Camera Similarity (IICS) and Intra- and Inter-Domain Similarity (IIDS) frameworks for unsupervised person re-identification (Re-ID). These frameworks follow a two-stage process that first generates pseudo labels by modeling both intra-camera and inter-camera relationships. These pseudo labels are then used to train feature encoders that learn identity representations consistent across multiple cameras. We improve upon this scheme by replacing the ResNet backbone with ConvNeXt—a convolutional architecture inspired by vision transformers, combining modern design principles with the efficiency of CNNs to achieve state-of-the-art performance in image recognition tasks. Additionally, we introduce two normalization techniques: (1) Adaptive Instance-Batch Normalization (AIBN) and (2) Transform Normalization (TNorm). Extensive ablation studies demonstrate that applying AIBN in the final ConvNeXt stages (Stages 3 and 4), and inserting TNorm after Stages 1 through 3, leads to significant performance improvements. We also analyze four ConvNeXt variants within the IICS/IIDS framework and demonstrate that larger ConvNeXt models consistently yield better performance. Experimental results on the Market1501, DukeMTMC-reID, and MSMT17 benchmarks show that our method achieves state-of-the-art performance among unsupervised person Re-ID approaches in terms of mean Average Precision (mAP), underscoring the potential of ConvNeXt-based architectures for scalable, label-free re-identification.

1 Introduction

Person Re-Identification (Re-ID) refers to the task of matching a person observed in one camera to the same individual seen earlier, either in the same or a different camera. This problem naturally arises in large-scale video surveillance systems and plays a critical role in applications such as security monitoring, search and rescue, and smart infrastructure. For example, in a shopping mall equipped with multiple cameras, person Re-ID enables operators to associate a visitor's current appearance with past sightings and to track their movement across non-overlapping camera views.

Despite a decade of research and growing interest from public safety agencies, person Re-ID remains an unsolved problem. The core challenge lies in reliably identifying individuals across diverse and uncontrolled conditions, such as variations in camera viewpoint, lighting, occlusion, posture, and clothing. These difficulties are amplified in real-world surveillance settings, where high-resolution imagery and biometric data are often unavailable, and captures may occur across different days or weeks.

In this work, we study the person Re-ID problem under practical assumptions: individuals are photographed by multiple cameras with non-overlapping views, and there are no clothing changes across captures. These assumptions are consistent with publicly available datasets—Market1501, DukeMTMC-reID, and MSMT17—where images of individuals are taken around the same time using multiple cameras. Consequently, clothing remains consistent, allowing us to focus on addressing challenges related to domain shift and feature discrimination.

A key requirement for person Re-ID is learning camera-invariant representations—features that capture identity while ignoring confounding factors such as background clutter, viewpoint, or style. Deep learning methods have become standard for this task, with CNN backbones (e.g., ResNet-50) pre-trained on ImageNet widely adopted as feature extractors. These features are typically matched using nearest-neighbor

[a] https://orcid.org/0000-0002-8992-3607

search or distance-based ranking. More recent methods integrate metric learning or domain adaptation to improve robustness across camera domains.

Among recent unsupervised Re-ID approaches, IICS and IIDS stand out for their two-stage pseudo-labeling strategies: intra-camera and inter-camera label generation (Xuan and Zhang, 2021; Xuan and Zhang, 2022). These methods also incorporate normalization strategies—Adaptive Instance and Batch Normalization (AIBN) and Transform Normalization (TNorm) into the ResNet backbone to mitigate intra- and inter-camera variations, respectively.

In this paper, we extend the IICS/IIDS framework by replacing the ResNet backbone with ConvNeXt, a convolutional architecture inspired by vision transformers. We evaluate the effectiveness of inserting AIBN and TNorm into different stages of ConvNeXt, and study their impact on the learned representations. Our findings show that AIBN improves intra-camera invariance when applied in the deeper stages (stages 3 and 4), while TNorm is most effective when used in early stages (stages 1–3), where it helps normalize style shifts across camera views. We further evaluate four ConvNeXt variants of increasing size and observe that larger models consistently yield higher mean Average Precision (mAP) scores. When AIBN and TNorm are properly integrated, our method achieves state-of-the-art performance on the Market1501 and MSMT17 datasets, and competitive results on DukeMTMC-reID.

2 Related Work

Person Re-Identification (Re-ID) is a critical problem in video surveillance, aiming to match individuals across non-overlapping camera views. Traditional approaches relied on hand-crafted features such as color histograms and texture descriptors (Liu et al., 2017; **?**), but the advent of deep learning, especially convolutional neural networks (CNNs), has led to a paradigm shift in Re-ID research (Li et al., 2014).

Supervised learning has driven rapid progress in Re-ID. Benchmarks such as Market1501 (Zheng et al., 2015) and DukeMTMC-reID (Ristani et al., 2016) show dramatic improvements, with Rank-1 accuracy increasing from $\tilde{4}0\%$ to over 95% in recent years (Liu et al., 2020). AlignedReID (Zhang et al., 2017) introduced a two-branch architecture combining global and local features with triplet hard loss. PCB (Part-based Convolutional Baseline) (Sun et al., 2018) further emphasized part-level representations with refined pooling techniques to capture discriminative regions. Luo et al. (Luo et al., 2019) proposed a strong baseline using ResNet50 with softmax and triplet loss, enhanced by the BNNeck design to separate metric and classification loss spaces. Other techniques include hash-code learning for efficient indexing (Liu et al., 2019), body-part attention for occlusion robustness (Somers et al., 2023), and clothing-invariant feature learning using adversarial loss (Gu et al., 2022). VP-ReID (Wei et al., 2018a) further explores retrieval acceleration using hierarchical clustering.

Due to the high cost of annotation, unsupervised approaches aim to learn Re-ID models from unlabeled data. These are commonly categorized into three groups. **(1)** Domain adaptation methods reduce feature gaps between source and target domains. CORAL (Sun and Saenko, 2016) aligns second-order statistics, while MMFA (Lin et al., 2018) aligns mid-level attribute features using MMD loss. Camera-aware similarity consistency learning (Wu et al., 2019a) improves feature coherence across views. AGD (Lu et al., 2022) uses geometric distillation and dreaming memory for incremental adaptation without data. **(1)** CycleGAN (Zhong et al., 2018b) enables image-to-image translation across camera domains using cycle consistency, facilitating camera-invariant representation learning. PTGAN (Wei et al., 2018b) preserves person identity while mapping style features. Advanced disentangling models separate id-related/unrelated features for improved transfer (Zou et al., 2020b). Such approaches often rely on adversarial and perceptual losses to preserve structure during translation (Zhu et al., 2017). **(3)** Pseudo-labeling techniques cluster feature representations to assign surrogate labels. PUL (Fan et al., 2018) adopts self-paced learning with k-means and CNN fine-tuning. BUC (Lin et al., 2019) uses bottom-up merging of clusters. MMCL (Wang and Zhang, 2020) combines memory banks with multi-label classification. Mutual-learning techniques such as NRMT (Zhao et al., 2020), MMT (Ge et al., 2020a), and MEB-Net (Zhai et al., 2020b) mitigate label noise through co-teaching or ensemble mechanisms. IICS (Xuan and Zhang, 2021) and IIDS (Xuan and Zhang, 2022) enhance label quality by performing intra- and inter-camera similarity analysis. They also introduce normalization layers (AIBN, TNorm) and use self-distillation to bridge domain gaps.

Commonly used benchmarks for evaluating Person Re-ID schemes include Market1501 (Zheng et al., 2015), DukeMTMC-reID (Ristani et al., 2016), and MSMT17 (Wei et al., 2018b). Market1501 includes 1,501 identities from 6 cameras, while Duke contains 702 identities from 8 cameras. MSMT17, the most challenging, features 4,101 identities from 15

indoor/outdoor cameras with significant variation in lighting and background conditions.

2.1 Contributions

While supervised Re-ID methods dominate performance metrics, their reliance on labeled data limits its scalability. Unsupervised approaches, especially those using pseudo-labels, offer practical alternatives. This work extends the IICS/IIDS framework by replacing ResNet with ConvNeXt (Liu et al., 2022), a modern CNN that achieves competitive performance with transformer-based models. We evaluate multiple ConvNeXt variants and explore the integration of AIBN and TNorm. The proposed scheme is evaluated on Market1501, DukeMTMC-reID, and MSMT17 benchmarks, and the experimental results demonstrate notable gains on Market1501 and MSMT17 datasets confirming the viability of modern CNNs in unsupervised Re-ID.

3 Methodology

Person Re-ID aims to retrieve the most visually similar instance of a query image $\mathbf{I}_q$ from a gallery set $\mathcal{G}$. This can be formulated as:

$$g = \underset{g \in [1,G]}{\arg\max}\, \text{sim}(\mathbf{I}_q, \mathbf{I}_g), \tag{1}$$

where $\mathbf{I}_g \in \mathcal{G}$ and $G = |\mathcal{G}|$. While some gallery images may have identity annotations, Re-ID is also useful in unlabelled settings such as public surveillance, where it supports person tracking across space and time.

The central component of any Re-ID pipeline is the feature extractor $\mathcal{F}$, which maps an input image to a feature vector in $\mathbb{R}^D$. The goal is to ensure that features of the same identity are closer together than those of different individuals. This leads to a distance-based formulation:

$$g = \underset{g \in [1,G]}{\arg\min}\, \|\mathcal{F}(\mathbf{I}_q) - \mathcal{F}(\mathbf{I}_g)\|^2, \tag{2}$$

where $\mathcal{F}(\mathbf{I};\Theta_e)$ denotes the feature extractor parameterized by Θ_e. Our objective is to learn a feature extractor that captures identity-discriminative features while being robust to variations such as lighting, viewpoint, occlusion, and camera styles—*without any labeled data*.

Images captured by any camera vary due to factors such as person identity, pose, orientation, and clothing. However, when images are captured across multiple cameras, additional variation arises from camera-specific artifacts like color response and placement. Work by Xuan and Zhang (Xuan and Zhang, 2021; Xuan and Zhang, 2022) argues convincingly that both intra-camera and inter-camera variations must be considered when designing a feature extractor for person re-identification (Re-ID). They propose a general framework that cleanly separates feature learning into two stages—inter- and intra-camera—to address these distinct sources of variation. Notably, their approach does not require labeled training data, making it particularly well-suited to real-world deployments, where acquiring large-scale labeled datasets is often infeasible. In this work, we adopt this two-stage framework as discussed below.

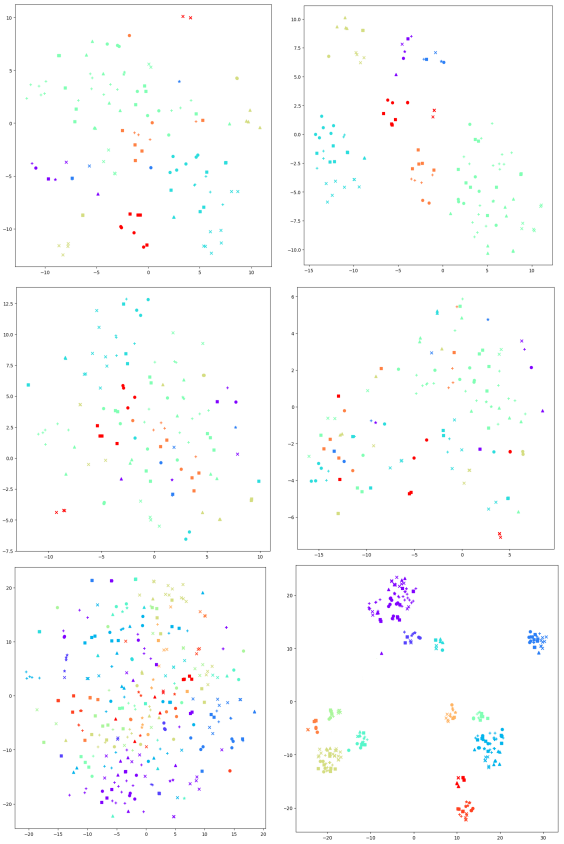

Figure 1: TSN-e plots visualizing the features computed using DukeMTMC-reID dataset. Here different markers represent different cameras and colors denote identities. (Top-row) the effects of intra-camera training, (middle-row) the effect of inter-camera training, and (bottom-row) the effect of inter- and intra-camera training.

3.1 Intra-Camera Training Stage

The *intra-camera stage* focuses on learning discriminative features within individual cameras. It consists of two steps: (A) camera-specific pseudo-label generation, and (B) feature extractor refinement.

3.1.1 Camera-Specific Pseudo-Labeling (A)

We begin by generating camera-specific pseudo-labels. Assume access to an initial feature extractor $\mathcal{F}$, commonly initialized using a model pre-trained on ImageNet (e.g., ResNet50). The training set $\mathcal{T} = \cup_{c=1}^{C} I_c$ includes images from C different cameras, where each I_c denotes the image set for camera c. For each camera c, we compute a set of features:

$$\mathcal{X}_c = \{\mathbf{x} = \mathcal{F}(\mathbf{I};\Theta_e) \mid \mathbf{I} \in I_c\}.$$

These features are clustered using *agglomerative hierarchical clustering* with average linkage, yielding clusters $\{\mathcal{P}_c^k\}_{k=1}^{K_c}$, such that $\mathcal{X}_c = \bigcup_{k=1}^{K_c} \mathcal{P}_c^k$ and $\mathcal{P}_c^i \cap \mathcal{P}_c^j = \varnothing$ for all $i \neq j$. Clustering is based on pairwise Euclidean distances in feature space.

Each image $\mathbf{I} \in I_c$ is then assigned a pseudo-label k if its feature belongs to cluster $\mathcal{P}_c^k$. These pseudo-labels are *camera-specific*, and thus labels across different cameras are not comparable—this distinction is addressed during the inter stage.

3.1.2 Feature Extractor Refinement (B)

Using the camera-specific pseudo-labels, we refine the feature extractor $\mathcal{F}$ through supervised learning. For each camera c, we define a K_c-way classifier $\mathcal{K}_c : \mathbf{x} \to \mathbb{R}^{K_c}$, parameterized by Θ_c. The joint parameters $\{\Theta_e, \Theta_1, \ldots, \Theta_C\}$ are optimized using the following loss:

$$l_{\text{intra}} = \sum_{\mathbf{I}\in\mathcal{T}} \mathbb{1}_{I_c}(\mathbf{I}) \cdot \text{cross-entropy}(\hat{\mathbf{p}}, \mathbf{p}),$$

where $\hat{\mathbf{p}} = \mathcal{K}_c(\mathcal{F}(\mathbf{I};\Theta_e);\Theta_c)$ and $\mathbf{p}$ is the one-hot encoding of the assigned pseudo-label for $\mathbf{I}$. The indicator function $\mathbb{1}_{I_c}(\mathbf{I})$ equals 1 if $\mathbf{I} \in I_c$ and 0 otherwise.

The cross-entropy loss is defined as:

$$\text{cross-entropy}(\hat{\mathbf{p}}, \mathbf{p}) = -\sum_i \mathbf{p}_i \log \hat{\mathbf{p}}_i.$$

This procedure trains both camera-specific classifiers and a shared feature extractor $\mathcal{F}$. The two steps—pseudo-labeling and refinement—can be repeated iteratively to progressively improve $\mathcal{F}$.

3.2 Inter-Camera Training Stage

The *inter stage* addresses variations across cameras. It mirrors the intra stage in structure, comprising (C) global pseudo-label generation and (D) feature extractor refinement.

3.2.1 Global Pseudo-Labeling (C)

Features for all training images $\mathcal{X} = \{\mathcal{F}(\mathbf{I};\Theta_e) \mid \mathbf{I} \in \mathcal{T}\}$ are clustered into K groups using agglomerative clustering with average linkage. In addition to Euclidean distance in feature space, clustering incorporates *Jaccard similarity* derived from camera-specific classifiers.

For a given image $\mathbf{I}$, each classifier $\mathcal{K}_c$ produces a K_c-dimensional probability vector. These outputs are concatenated and normalized into a distribution $\mathbf{q}$. The Jaccard similarity between two images $\mathbf{I}_l$ and $\mathbf{I}_m$ is computed as:

$$\Delta(\mathbf{I}_l, \mathbf{I}_m) = \frac{|\mathbf{q}_l \cap \mathbf{q}_m|}{|\mathbf{q}_l \cup \mathbf{q}_m|}.$$

Each image is assigned to a global cluster $\mathcal{P}^k$ based on its feature, resulting in cross-camera pseudo-labels.

3.2.2 Feature Extractor Refinement (D)

Using the global pseudo-labels, we refine $\mathcal{F}$ via supervised training with a K-way classifier $\mathcal{K} : \mathbf{x} \to \mathbb{R}^K$, parameterized by Θ. The optimization minimizes:

$$l_{\text{inter}} = \sum_{\mathbf{I}\in\mathcal{T}} \text{cross-entropy}(\hat{\mathbf{p}}, \mathbf{p}),$$

where $\hat{\mathbf{p}} = \mathcal{K}(\mathcal{F}(\mathbf{I};\Theta_e);\Theta)$ is the predicted label and $\mathbf{p}$ is the one-hot encoded pseudo-label.

As in the intra stage, this process can be repeated to further refine the model.

3.3 Overall Framework

The complete framework proposed in (Xuan and Zhang, 2021; Xuan and Zhang, 2022) alternates between intra- and inter-stage training. Specifically, the following regime is adopted:

$$[(A,B) \times 3 \text{ followed by } (C,D) \times 2] \times 40.$$

We will soon see that both intra- and inter-stage learning are essential. Omitting either leads to a significant drop in performance, highlighting the importance of jointly addressing within-camera and cross-camera variations.

3.4 Feature Extractor: ConvNeXt

Unlike the original IICS/IIDS which use ResNet50, we replace the backbone with ConvNeXt (Liu et al., 2022)—a modern CNN architecture inspired by vision transformers. We integrate AIBN and TNorm into different ConvNeXt stages to improve domain robustness. In our experiments, placing AIBN in stages 3 and 4, and TNorm in stages 1–3 yields the best results. For details about ConvNeXT, we refer the kind reader to (Liu et al., 2022).

Table 1: Results on MSMT17 dataset. Pseudo Label* indicates that the psuedo labels are initially constructed by using a pre-trained person Re-ID model. DA denotes Distribution Alignment.

Type	Method (Reference & Venue)	MSMT17				
		Source	mAP	Rank-1	Rank-5	Rank-10
GANs	PTGAN (Wei et al., 2018b) (CVPR 2018)	Market	2.9	10.2	-	24.4
	ECN (Zhong et al., 2019) (CVPR 2019)	Market	8.5	25.3	36.3	42.1
	SSG (Fu et al., 2019) (ICCV 2019)	Market	**13.2**	**31.6**	-	**49.6**
DA	NRMT (Zhao et al., 2020) (ECCV 2020)	Market	19.8	43.7	56.5	62.2
	DG-Net++ (Zou et al., 2020a) (ECCV 2020)	Market	22.1	48.4	60.9	66.1
	MMT-1500 (Ge et al., 2020a) (ICLR 2020)	Market	**22.9**	**49.2**	**63.1**	**68.8**
Pseudo Label*	PTGAN (Wei et al., 2018b) (CVPR 2018)	Duke	3.3	11.8	-	27.4
	ECN (Zhong et al., 2019) (CVPR 2019)	Duke	10.2	30.2	41.5	46.8
	SSG (Fu et al., 2019) (ICCV 2019)	Duke	13.3	32.2	-	51.2
	NRMT (Zhao et al., 2020) (ECCV 2020)	Duke	20.6	45.2	57.8	63.3
	DG-Net++ (Zou et al., 2020a) (ECCV 2020)	Duke	22.1	48.8	60.9	65.9
	MMT-1500 (Ge et al., 2020a) (ICLR 2020)	Duke	**23.3**	**50.1**	**63.9**	**69.8**
Pseudo Label	MMCL (Wang and Zhang, 2020) (CVPR 2020)	None	11.2	35.4	44.8	49.8
	JVTC+ (Zhang et al., 2021) (ECCV 2020)	None	17.3	43.1	53.8	59.4
	SpCL (Ge et al., 2020b) (NeurIPS 2020)	None	19.1	42.3	55.6	61.2
	IICS (Xuan and Zhang, 2021) (CVPR 2021)	None	26.9	56.4	68.8	73.4
	IIDS (Xuan and Zhang, 2022) (CVPR 2022)	None	**35.1**	**64.4**	**76.2**	**80.5**
Our Method	Iso-ConvNeXt-S	None	27.5	57.3	69.1	74.5
	Iso-ConvNeXt-S (AIBN)	None	29.6	60.0	72.4	77.9
	Iso-ConvNeXt-S (AIBN, TNorm)	None	36.4	65.1	77.8	82.6
	ConvNeXt-B (AIBN, TNorm)	None	**40.2**	**71.3**	**82.0**	**86.3**

3.5 Adaptive Instance and Batch Normalization

Adaptive Instance and Batch Normalization (AIBN) builds on Batch-Instance Normalization (BIN) by adaptively blending Batch Normalization and Instance Normalization through learnable gates, allowing the model to balance global statistics with per-sample style adjustments. In ConvNeXt-based person re-identification, AIBN is applied in later stages (Stages 3 and 4) to effectively normalize camera-specific variations while preserving identity-discriminative features, leading to improved mAP on benchmarks like Market1501 and MSMT17.

Mathematically, let $\mathbf{x} \in \mathbb{R}^{N\times C\times H\times W}$ be the input tensor. Here N denotes the batch size. Then Batch Normalization (BN) statistics are

$$\mu^c_{\text{bn}} = \frac{1}{NHW}\sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{x}_{nchw}, \text{ and}$$
$$\sigma^{2,c}_{\text{bn}} = \frac{1}{NHW}\sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W}(\mathbf{x}_{nchw}-\mu^c_{\text{bn}})^2.$$

Similarly, Instance Normalization (IN) statistics are

$$\mu^{n,c}_{\text{in}} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{x}_{nchw}, \text{ and}$$
$$\sigma^{2,n,c}_{\text{in}} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(\mathbf{x}_{nchw}-\mu^{n,c}_{\text{in}})^2.$$

AIBN statistics with learnable gate $\rho^c \in [0,1]$ are

$$\mu^{n,c}_{\text{aibn}} = \rho^c\cdot\mu^{n,c}_{\text{in}} + (1-\rho^c)\cdot\mu^c_{\text{bn}}, \text{ and}$$
$$\sigma^{n,c}_{\text{aibn}} = \rho^c\cdot\sigma^{n,c}_{\text{in}} + (1-\rho^c)\cdot\sigma^c_{\text{bn}}.$$

This yields

$$\mathbf{y}_{nchw} = \gamma^c\cdot\hat{\mathbf{x}}_{nchw} + \beta^c, \text{where}$$
$$\hat{\mathbf{x}}_{nchw} = \frac{\mathbf{x}_{nchw}-\mu^{n,c}_{\text{aibn}}}{\sqrt{\sigma^{n,c}_{\text{aibn}}+\varepsilon}}.$$

3.6 Transform Normalization

Transform Normalization (TNorm) is a lightweight, learnable normalization layer designed to mitigate domain shifts in unsupervised person re-identification by applying a camera-aware linear transformation followed by normalization. Introduced in the MCRN framework (Zhang et al., 2022), TNorm is particularly effective in early network stages where low-level variations are prominent, complementing deeper-layer techniques like AIBN. When integrated into architectures like ResNet or ConvNeXt, TNorm improves performance across datasets such as Market1501 and MSMT17 by enabling better camera-specific feature alignment without requiring labels.

Mathematically, let $\mathbf{x} \in \mathbb{R}^{N\times C}$ be a feature vector from the output of a backbone network, and let $c \in \{1,\ldots,C'\}$ denote the camera index (domain). For each camera c, Transform Normalization (TNorm)

Table 2: Results on Market1501 dataset. Pseudo Label* indicates that the psuedo labels are initially constructed by using a pre-trained person Re-ID model. DA stands for Distribution Alignment.

Type	Method (Reference & Venue)	Market1501				
		Source	mAP	Rank-1	Rank-5	Rank-10
GANs	PTGAN (Wei et al., 2018b) (CVPR 2018)	Duke	-	38.6	-	66.1
	HHL (Zhong et al., 2018a) (ECCV 2018)	Duke	31.4	62.2	78.8	84.0
	DG-Net++ (Zou et al., 2020a) (ECCV 2020)	Duke	**61.7**	**82.1**	**90.2**	**92.7**
DA	TJ-AIDL (Wang et al., 2018) (CVPR 2018)	Duke	26.5	58.2	74.8	81.8
	MMFA (Lin et al., 2018) (BMVC 2018)	Duke	27.4	56.7	75.0	81.8
	CSCL (Wu et al., 2019b) (ICCV 2019)	Duke	**35.6**	**64.7**	**80.2**	**85.6**
Pseudo Label*	MAR (Yu et al., 2019) (CVPR 2019)	MSMT17	40.0	67.7	81.9	-
	AD-Cluster (Zhai et al., 2020a) (CVPR 2020)	Duke	68.3	86.7	94.4	96.5
	NRMT (Zhao et al., 2020) (ECCV 2020)	Duke	71.7	87.8	94.6	96.5
	MMT-500 (Ge et al., 2020a) (ICLR 2020)	Duke	71.2	87.7	94.9	**96.9**
	MEB-Net* (Zhai et al., 2020b) (ECCV 2020)	Duke	**71.9**	**87.5**	**95.2**	96.8
Pseudo Label	LOMO (Liao et al., 2015) (CVPR 2015)	None	8.0	27.2	41.6	49.1
	BOW (Zheng et al., 2015) (ICCV 2015)	None	14.8	35.8	52.4	60.3
	BUC (Lin et al., 2019) (AAAI 2019)	None	29.6	61.9	73.5	78.2
	HCT (Zeng et al., 2020) (CVPR 2020)	None	56.4	80.0	91.6	95.2
	MMCL (Wang and Zhang, 2020) (CVPR 2020)	None	45.5	80.3	89.4	92.3
	JVTC+ (Li and Zhang, 2020) (ECCV 220)	None	47.5	79.5	89.2	91.9
	IICS (Xuan and Zhang, 2021) (CVPR 2021)	None	72.1	88.8	95.3	96.9
	IIDS (Xuan and Zhang, 2022) (CVPR 2022)	None	**78.3**	**91.2**	**96.2**	**97.7**
Our Method	Iso-ConvNeXt-S	None	72.7	89.8	95.4	97.2
	Iso-ConvNeXt-S (AIBN)	None	74.8	91.4	97.2	98.0
	Iso-ConvNeXt-S (AIBN, TNorm)	None	79.7	94.6	98.1	98.7
	ConvNeXt-B (AIBN, TNorm)	None	**83.1**	**97**	**99.2**	**99.6**

computes camera-specific normalization as:

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i - \mu^{(c)}}{\sqrt{\sigma^{2(c)} + \varepsilon}},$$

where

$$\mu^{(c)} = \frac{1}{N_c} \sum_{i \in I_c} \mathbf{x}_i, \text{ and}$$

$$\sigma^{2(c)} = \frac{1}{N_c} \sum_{i \in I_c} (\mathbf{x}_i - \mu^{(c)})^2.$$

Here I_c denotes the set of samples from camera c, and $N_c = |I_c|$. The Affine transformation is shared across all camera

$$\mathbf{y}_i = \gamma \cdot \hat{\mathbf{x}}_i + \beta,$$

where γ and β are learnable parameters shared across all cameras, and ε is a small constant for numerical stability.

3.7 The Need for Intra- and Inter-Camera Training Stages

We investigate the effect of intra- and inter-camera training on feature representation. Figure 1 presents t-SNE visualizations of features extracted from multiple identities across different cameras. Each marker shape corresponds to a specific camera, while colors represent different identities. The left column shows the feature distributions before training, and the right column shows them after training. The top row corresponds to inter-camera training only, the middle row to intra-camera training only (note that Jaccard similarity is unavailable in this case, as it becomes accessible only after inter-camera training), and the bottom row corresponds to both intra- and inter-camera training stages.

An ideal outcome is one where features of the same identity—despite being captured by different cameras—cluster together in the embedding space. As the plots demonstrate, neither inter-camera nor intra-camera training alone is sufficient to achieve this. Only the joint training setup results in identity-consistent clustering, as clearly illustrated in Figure 1 (bottom-right).

4 Results

We evaluated our method against other unsupervised and transfer learning approaches on three widely-used datasets: Market1501, DukeMTMC-reID, and MSMT17. Tables 1, 2, and 3 present comparisons of different variants of our method with existing approaches on these datasets, respectively. All reported accuracy values for our model are averaged over three runs using different random seeds to ensure robustness.

Table 3: Results on DukeMTMC dataset. Pseudo Label* indicates that the psuedo labels are initially constructed by using a pre-trained person Re-ID model. DA stands for Distribution Alignment.

Type	Method (Reference & Venue)	DukeMTMC-reID				
		Source	mAP	Rank-1	Rank-5	Rank-10
GANs	PTGAN (Wei et al., 2018b) (CVPR 2018)	Market	-	27.4	-	50.7
	HHL (Zhong et al., 2018a) (ECCV 2018)	Market	27.2	46.9	61.0	66.7
	DG-Net++ (Zou et al., 2020a) (ECCV 2020)	Market	**63.8**	**78.9**	**87.8**	**90.4**
DA	TJ-AIDL (Wang et al., 2018) (CVPR 2018)	Market	23.0	44.3	59.6	65.0
	MMFA (Lin et al., 2018) (BMVC 2018)	Market	24.7	45.3	59.8	66.3
	CSCL (Wu et al., 2019b) (ICCV 2019)	Market	**30.5**	**51.5**	**66.7**	**71.7**
Pseudo Label*	MAR (Yu et al., 2019) (CVPR 2019)	MSMT17	48.0	67.1	79.8	-
	AD-Cluster (Zhai et al., 2020a) (CVPR 2020)	Market	54.1	72.6	82.5	85.5
	NRMT (Zhao et al., 2020) (ECCV 2020)	Market	62.2	77.8	86.9	89.5
	MMT-500 (Ge et al., 2020a) (ICLR 2020)	Market	63.1	76.8	**88.0**	**92.2**
	MEB-Net* (Zhai et al., 2020b) (ECCV 2020)	Market	**63.5**	**77.2**	87.9	91.3
Pseudo Label	LOMO (Liao et al., 2015) (CVPR 2015)	None	4.8	12.3	21.3	26.6
	BOW (Zheng et al., 2015) (ICCV 2015)	None	8.3	17.1	28.8	34.9
	BUC (Lin et al., 2019) (AAAI 2019)	None	22.1	40.4	52.5	58.2
	HCT (Zeng et al., 2020) (CVPR 2020)	None	50.7	69.6	83.4	87.4
	MMCL (Wang and Zhang, 2020) (CVPR 2020)	None	40.2	65.2	75.9	80.0
	JVTC+ (Li and Zhang, 2020) (ECCV 2020)	None	50.7	74.6	82.9	85.3
	IICS (Xuan and Zhang, 2021) (CVPR 2021)	None	59.1	76.9	86.1	89.8
	IIDS (Xuan and Zhang, 2022) (CVPR 2022)	None	**68.7**	**82.1**	**90.8**	**93.7**
Our Method	Iso-ConvNeXt-S	None	48.2	67.1	77.3	80.6
	Iso-ConvNeXt-S (AIBN)	None	54.3	72.8	81.3	84.6
	Iso-ConvNeXt-S (AIBN, TNorm)	None	60.8	78.3	85.5	89.8
	ConvNeXt-B (AIBN, TNorm)	None	**65.2**	**80.3**	**83.4**	**87.6**

The last four rows of each table detail the performance of our method in various configurations. Iso-ConvNeXt-S refers to the smallest isotropic version of ConvNeXt, used as the feature extractor within the IICS/IIDS framework. Iso-ConvNeXt-S (AIBN) incorporates AIBN into the ConvNeXt backbone, while Iso-ConvNeXt-S (AIBN, TNorm) includes both AIBN and TNorm. The best performance is achieved by ConvNeXt-B (AIBN, TNorm), a larger ConvNeXt variant enhanced with both normalization techniques.

Our method is compared with several state-of-the-art approaches, including GAN-based models such as PTGAN (Wei et al., 2018b), distribution alignment-based methods like TJ-AIDL (Wang et al., 2018), and pseudo-labeling strategies such as MAR (Yu et al., 2019). Among these, pseudo-label-based techniques consistently outperform others.

Results in tables 1, 2, and 3 demonstrate that our approach surpasses both IICS and IIDS methods on the MSMT17 and Market1501 datasets. The performance is comparable to (Xuan and Zhang, 2022) on Duke dataset. We suggest the reader to pay particular attention to Iso-ConvNext-S (AIBN, TNorm) configuration, which, roughly speaking, has the same number of parameters as the ResNet50 backbone used in IICS/IIDS (see Table 5). Table 4 compares the performance for ConvNeXT variants used in this work. The improved performance of ConvNeXt over the ResNet backbone—originally used in IICS/IIDS—can be attributed to its use of depthwise and pointwise convolutions. This architectural design separates spatial and channel-wise processing, enabling more effective learning of discriminative spatial features (e.g., body structure) and appearance cues (e.g., clothing). Such separation enhances the model's capacity to distinguish individuals, leading to better person re-identification.

4.1 Ablation study

Intra- and inter-camera training stages: We investigate the impact of intra-camera and inter-camera training. In the first experiment, we use a pre-trained ConvNeXt model without any fine-tuning. In subsequent experiments, we examine the effects of applying only intra-camera or only inter-camera training. For the inter-camera training stage, we compute feature similarity using CNN embeddings, without incorporating Jaccard similarity. Finally, we perform both intra- and inter-camera training using ConvNeXt as the feature extractor, without incorporating AIBN or TNorm enhancements. As shown in Table 6, the highest performance is achieved when both intra- and inter-camera training are applied.

Effectiveness of AIBN and TNorm, and Their Insertion Locations: To address intra-camera and inter-camera variations during feature extraction, we integrate two normalization techniques—AIBN and

Table 4: Variants of ConvNeXt used in this paper. Models were pre-trained on ImageNet1K.

Dataset		Market		Duke		MSMT	
Set up		mAP	Rank1	mAP	Rank1	mAP	Rank1
Iso-ConvNeXt-S	AIBN	74.8	91.4	54.3	72.8	29.6	60.0
	AIBN+TNorm	79.7	94.6	60.8	78.3	36.4	65.1
ConvNeXt-T	AIBN	75.2	92.3	55.1	73.5	30.2	61.4
	AIBN+TNorm	81.5	95.6	62.0	91.3	38.6	68.3
ConvNeXt-S	AIBN	75.9	93.0	55.7	74.1	32.5	62.7
	AIBN+TNorm	82.3	96.1	62.5	**91.9**	39.1	69.4
ConvNeXt-B	AIBN	76.5	93.8	56.5	75.3	34.0	63.5
	AIBN+TNorm	**83.1**	**97**	**65.2**	80.3	**40.2**	**71.3**

Table 5: Model complexity and parameter counts of various architectures. 'M' denotes millions of parameters, and 'G' denotes computational complexity in gigaflops (GFLOPs), following the convention in (Liu et al., 2022).

Architecture	#Params	FLOPs
ResNet-50	25.6M	4.1G
Iso-ConvNeXt-S	22M	4.3G
ConvNeXt-T	28.6M	4.5G
ConvNeXt-S	49.6M	8.7G
ConvNeXt-B	88.6M	15.4G

Table 6: The role of inter- and intra-camera training stages.

Dataset	Market		Duke	
	mAP	Rank1	mAP	Rank1
Pretrain	5.7	16.5	4.8	13.2
Intra	46.3	69.9	28.1	45.8
Inter (w/o Jaccard)	27.2	48.8	8.2	17.1
Intra + Inter	**72.7**	**89.8**	**48.2**	**67.1**

TNorm—into the ConvNeXt architecture. AIBN is designed to reduce intra-camera variations arising from differences in pose, appearance, and other identity-specific factors. In contrast, TNorm targets inter-camera variations caused by differences in camera characteristics such as color shifts. Our experiments show that replacing all LayerNorm (LN) layers in ConvNeXt with AIBN leads to a drop in performance. However, selectively applying AIBN in stages 3 and 4 improves results. The detailed impact of different AIBN insertion points is summarized in Table 7. To further mitigate inter-camera discrepancies, we insert TNorm layers after specific stages of ConvNeXt. As shown in Table 8, the best performance is achieved when TNorm is applied after stages 1, 2, and 3.

5 Conclusions

This work enhances unsupervised person re-identification by integrating ConvNeXt into the IICS/IIDS framework. Replacing ResNet with ConvNeXt, especially larger variants, improves accuracy. We introduce two key normalization strategies—AIBN and TNorm—and show that their strategic placement (AIBN in final stages, TNorm after early stages) significantly boosts performance. Combining intra- and inter-camera training further strengthens identity consistency. Experiments on Market1501, DukeMTMC, and MSMT17 confirm our method outperforms recent IICS/IIDS variants, particularly on Market1501 and MSMT17.

5.1 Ethical and Societal Concerns

Person Re-ID systems raise significant ethical and societal concerns, primarily due to their potential to infringe on individual privacy by enabling continuous, unconsented surveillance across public and private spaces. These systems often operate without transparency or accountability, making it difficult to contest or audit misidentifications, particularly when powered by opaque deep learning models. Moreover, Re-ID technologies may exhibit bias against underrepresented demographic groups due to imbalanced training data, leading to unfair treatment or disproportionate surveillance of marginalized communities. The widespread deployment of such systems risks normalizing constant monitoring, eroding public trust and potentially enabling function creep—where technologies initially intended for safety are repurposed for social control, commercial profiling, or political suppression. Addressing these challenges requires not only technical safeguards such as fairness-aware training and explainability, but also robust legal and ethical frameworks to ensure responsible use.

5.2 The Use of DukeMTMC-reID

Despite its known privacy concerns and the fact that it was decommissioned due to lack of consent in the original data collection, the DukeMTMC-reID dataset is still used in academic settings because of its challenging multi-camera setup, rich annotations, and its status as a historical benchmark that enables fair comparisons with prior work. For the sake of full-disclosure, we decided to include the results on this

Table 7: Ablation study evaluating the impact of inserting AIBN at different stages of the ConvNeXt architecture. For example, "AIBN (Block 2-3-4)" indicates that AIBN replaces the LayerNorm layers in stages 2 through 4. The best performance is achieved when AIBN is applied only in the final two stages (stages 3 and 4).

Dataset	**Market**		**Duke**	
Setup	mAP	Rank1	mAP	Rank1
Baseline	72.7	89.8	48.2	67.1
All	50.3	69.9	28.1	45.8
AIBN (Block 1-2)	69.2	85.3	35.6	58.2
AIBN (Block 4)	73.2	90	51.8	71.1
AIBN (Block 2-3-4)	74.2	91.2	53.9	72.3
AIBN (Block 3-4)	**74.8**	**91.4**	**54.4**	**72.8**

Table 8: Ablation study on the effect of inserting TNorm at different stages of the ConvNeXt architecture. For instance, "TNorm (stage 1-2-3-4)" indicates that TNorm is inserted after stages 1 through 4. The highest performance is observed when TNorm is applied after stages 1, 2, and 3.

Dataset	**Market**		**Duke**	
Setup	mAP	Rank1	mAP	Rank1
Baseline	72.7	89.8	48.2	67.1
TNorm (Stage 1)	74.3	92.1	50.9	69.2
TNorm (Stage 1-2)	74.9	92.7	51.8	71.1
TNorm (Stage 1-2-3)	**75.4**	**92.4**	**57.8**	**76.3**
TNorm (Satge 1-2-3-4)	74.2	92	50.2	70.1

dataset despite the fact that our model did not achieve state-of-the-art mAP scores on this dataset.

5.3 Limitations and Final Word

Despite these gains, our approach faces limitations in scalability, demographic generalization, and robustness to occlusion and crowded scenes. Future work will explore stronger occlusion modeling, domain adaptation for broader demographics, and privacy-aware dataset construction. Our findings open promising directions for designing scalable, adaptive, and robust Re-ID systems under real-world, camera-diverse, and label-sparse conditions.

ACKNOWLEDGEMENTS

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), RGPIN-2020-05159.

REFERENCES

Fan, H., Zheng, L., Yan, C., and Yang, Y. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–18. 2

Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., and Huang, T. S. (2019). Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6112–6121, Seoul, South Korea. 5

Ge, Y., Chen, D., and Li, H. (2020a). Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*. 2, 5, 6, 7

Ge, Y., Zhu, F., Chen, D., Zhao, R., et al. (2020b). Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in neural information processing systems*, 33:11309–11321. 5

Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., and Chen, X. (2022). Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1060–1069, New Orleans, LA, USA. IEEE/CVF. 2

Li, J. and Zhang, S. (2020). Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 483–499. Springer. 6, 7

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR, Columbus, Ohio, USA*. 2

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206, Boston, MA, USA. 6, 7

Lin, S., Li, H., Li, C.-T., and Kot, A. C. (2018). Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint arXiv:1807.01440*. 2, 6, 7

Lin, Y., Dong, X., Zheng, L., Yan, Y., and Yang, Y. (2019). A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8738–8745. 2, 6, 7

Liu, C., Chang, X., and Shen, Y.-D. (2020). Unity style transfer for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6887–6896. 2

Liu, W., Chang, X., Chen, L., and Yang, Y. (2017). Early active learning with pairwise constraint for person re-identification. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pages 103–118. Springer. 2

Liu, X., Zhang, S., and Yang, M. (2019). Self-guided hash coding for large-scale person re-identification. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 246–251. IEEE. 2

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986. 3, 4, 8

Lu, Y., Wang, M., and Deng, W. (2022). Augmented geometric distillation for data-free incremental person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7329–7338, New Orleans, LA, USA. IEEE/CVF. 2

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., and Gu, J. (2019). A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609. 2

Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 17–35. Springer. 2

Somers, V., De Vleeschouwer, C., and Alahi, A. (2023). Body part-based representation learning for occluded person re-identification. In *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1613–1623, Waikoloa, HI, USA. IEEE/CVF. 2

Sun, B. and Saenko, K. (2016). Deep CORAL: correlation alignment for deep domain adaptation. *CoRR*, abs/1607.01719. 2

Sun, Y., Zheng, L., Yang, Y., Tian, Q., and Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, Munich, Germany. 2

Wang, D. and Zhang, S. (2020). Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10981–10990. 2, 5, 6, 7

Wang, J., Zhu, X., Gong, S., and Li, W. (2018). Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2275–2284, Salt Lake City, UT, USA. 6, 7

Wei, L., Liu, X., Li, J., and Zhang, S. (2018a). Vp-reid: Vehicle and person re-identification system. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 501–504. 2

Wei, L., Zhang, S., Gao, W., and Tian, Q. (2018b). Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 79–88, Salt Lake City, UT, USA. 2, 5, 6, 7

Wu, A., Zheng, W.-S., and Lai, J. (2019a). Unsupervised person re-identification by camera-aware similarity consistency learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6921–6930. 2

Wu, A., Zheng, W.-S., and Lai, J.-H. (2019b). Unsupervised person re-identification by camera-aware similarity consistency learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6922–6931. 6, 7

Xuan, S. and Zhang, S. (2021). Intra-inter camera similarity for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11926–11935. 2, 3, 4, 5, 6, 7

Xuan, S. and Zhang, S. (2022). Intra-inter domain similarity for unsupervised person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2, 3, 4, 5, 6, 7

Yu, H.-X., Zheng, W.-S., Wu, A., Guo, X., Gong, S., and Lai, J.-H. (2019). Unsupervised person re-identification by soft multilabel learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2148–2157. 6, 7

Zeng, K., Ning, M., Wang, Y., and Guo, Y. (2020). Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13657–13665, Virtual Event. 6, 7

Zhai, Y., Lu, S., Ye, Q., Shan, X., Chen, J., Ji, R., and Tian, Y. (2020a). Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9021–9030. 6, 7

Zhai, Y., Ye, Q., Lu, S., Jia, M., Ji, R., and Tian, Y. (2020b). Multiple expert brainstorming for domain adaptive person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII*, pages 594–611. Springer. 2, 6, 7

Zhang, K., Xu, H., Wang, Z., Wang, L., Lin, W., and Chua, T.-S. (2022). Unsupervised person re-identification via meta camera style adaptation. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, volume 36, pages 1285–1293. 5

Zhang, T., Xie, L., Wei, L., Zhuang, Z., Zhang, Y., Li, B., and Tian, Q. (2021). Unrealperson: An adaptive pipeline towards costless person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11506–11515, Nashville, TN, USA. 5

Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., and Sun, J. (2017). Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*. 2

Zhao, F., Liao, S., Xie, G.-S., Zhao, J., Zhang, K., and Shao, L. (2020). Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 526–544. Springer. 2, 5, 6, 7

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, Santiago, Chile. 2, 6, 7

Zhong, Z., Zheng, L., Li, S., and Yang, Y. (2018a). Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–188, Munich, Germany. 6, 7

Zhong, Z., Zheng, L., Luo, Z., Li, S., and Yang, Y. (2019). Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 598–607, Long Beach, CA, USA. 5

Zhong, Z., Zheng, L., Zheng, Z., Li, S., and Yang, Y. (2018b). Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5157–5166, Salt Lake City, UT, USA. 2

Zhu, J., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593. 2

Zou, Y., Yang, X., Yu, Z., Kumar, B. V., and Kautz, J. (2020a). Joint disentangling and adaptation for cross-domain person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 87–104. Springer. 5, 6, 7

Zou, Y., Yang, X., Yu, Z., Kumar, B. V. K. V., and Kautz, J. (2020b). Joint disentangling and adaptation for cross-domain person re-identification. *CoRR*, abs/2007.10315. 2