Towards Anytime Retrieval: A Benchmark for Anytime Person Re-Identification

Xulin Li[1,2], **Yan Lu**[3], **Bin Liu**[1,2*], **Jiaze Li**[1,2], **Qinhong Yang**[1,2], **Tao Gong**[1,2], **Qi Chu**[1,2], **Mang Ye**[4] and **Nenghai Yu**[1,2]

[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]Anhui Province Key Laboratory of Digital Security
[3]The Chinese University of Hong Kong
[4]School of Computer Science, Wuhan University, China
lxlkw@mail.ustc.edu.cn, yanlu@cuhk.edu.hk, flowice@ustc.edu.cn, jz_li@mail.ustc.edu.cn, qhyang233@mail.ustc.edu.cn, {tgong,qchu}@ustc.edu.cn, yemang@whu.edu.cn, ynh@ustc.edu.cn

Abstract

In real applications, person re-identification (ReID) expects to retrieve the target person at any time, including both daytime and nighttime, ranging from short-term to long-term. However, existing ReID tasks and datasets cannot meet this requirement, as they are constrained by available time and only provide training and evaluation for specific scenarios. Therefore, we investigate a new task called Anytime Person Re-identification (AT-ReID), which aims to achieve effective retrieval in multiple scenarios based on variations in time. To address the AT-ReID problem, we collect the first large-scale dataset, AT-USTC, which contains 135k images of individuals wearing multiple clothes captured by RGB and IR cameras. Our data collection spans over an entire year and 270 volunteers were photographed on average 29.1 times across different dates or scenes, 4-15 times more than current datasets, providing conditions for follow-up investigations in AT-ReID. Further, to tackle the new challenge of multi-scenario retrieval, we propose a unified model named Uni-AT, which comprises a multi-scenario ReID (MS-ReID) framework for scenario-specific features learning, a Mixture-of-Attribute-Experts (MoAE) module to alleviate inter-scenario interference, and a Hierarchical Dynamic Weighting (HDW) strategy to ensure balanced training across all scenarios. Extensive experiments show that our model leads to satisfactory results and exhibits excellent generalization to all scenarios.

1 Introduction

Person re-identification (ReID) aims to retrieve specific pedestrians with given query images. As illustrated in Fig. 1 (a), a robust ReID system is expected to retrieve a person at any time, including daytime and nighttime, ranging from short-term to long-term, thereby satisfying the requirements of different surveillance scenarios. This puts more challenges on the ReID system because the capturing time of the query image and the target image makes the task more variable. For instance, if two images are captured during daytime and nighttime, respectively, they will have different modalities, and when there is a long time interval between their capturing, the person's appearance may change due to alterations in clothing. Consequently, traditional ReID (Tr-ReID) [Zheng *et al.*, 2015] may not perform effectively.

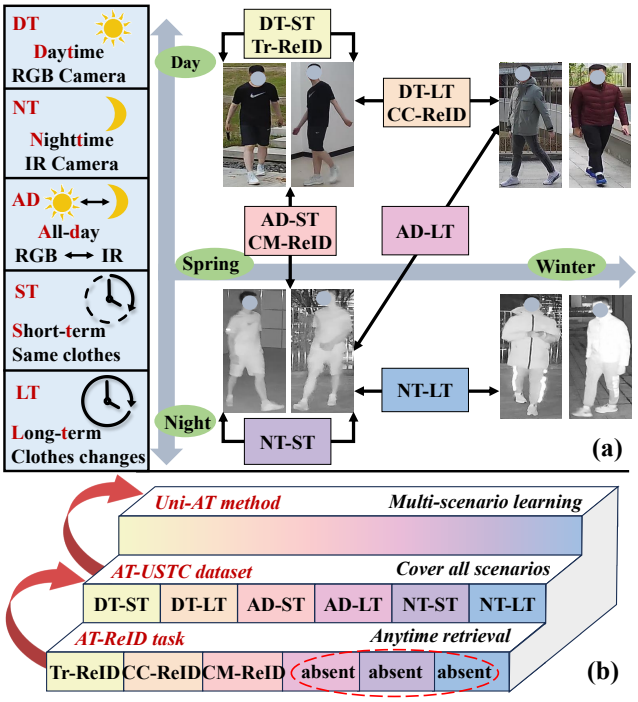

Figure 1: (a) Six non-overlapping scenarios based on variations in time. AT-ReID aims to perform retrieval in all of these scenarios. (b) Our solution of AT-ReID from the dataset to the model level.

The researchers acknowledged this challenge and attempted to address these problems separately. They introduced the Cross-Modality ReID (CM-ReID) [Wu *et al.*, 2017] to address the issue of searching between daytime RGB images and nighttime infrared (IR) images, and the Cloth-

*Corresponding author.

Changing ReID (CC-ReID) [Yang *et al.*, 2019] was proposed to handle long-term retrieval in which pedestrians change their clothes. However, existing methods designed for these specific tasks were only able to achieve success in one of them and incapable of retrieving targets at any time simultaneously. This situation primarily arises from the absence of a long-term visible-infrared dataset covering all scenarios in Fig. 1 (a), which should encompass diverse variations in clothing and modality for each individual. As shown in Fig. 1 (b), the deficiency in intra-identity diversity of modalities or clothing in current ReID datasets has led to research gaps, especially in Nighttime Long-term (NT-LT) and All-day Long-term (AD-LT) scenarios. Another issue arises from the poor generalization of task-specific methods in non-target scenarios. This is attributed to the differing learning objectives across different scenarios. For instance, prior research [Lu *et al.*, 2020; Qian *et al.*, 2020] has indicated that RGB-specific cues and clothing information are harmful to the All-day Short-term (AD-ST, CM-ReID) and Daytime Long-term (DT-LT, CC-ReID) scenarios, respectively, while they are crucial for the Daytime Short-term (DT-ST, Tr-ReID) scenario.

To meet the requirements of retrieving persons at any time, we investigate a new task called **Anytime Person Re-identification (AT-ReID)** and propose to focus on its exploration from dataset to model level. We collect the first corresponding large-scale dataset named **AT-USTC**[1], which contains 135k images and covers all six scenarios in AT-ReID. Our data collection spans an entire year, covering both day and night periods across the seasons of spring, summer, and winter. We focus on simultaneously providing a greater variety of clothing and more RGB and IR cameras for each person. Through efforts to expand in terms of capture dates, time periods, and scene variations, our AT-USTC provides a broader intra-identity diversity and more comprehensive AT-ReID cases than previous datasets.

To tackle the new challenge of multi-scenario retrieval in AT-ReID, we further propose a unified model named **Uni-AT** to effectively handle all scenarios. Given that the AT-ReID encompasses six different scenarios, the information shared among all scenarios becomes limited, and learning a unified representation for all scenarios is sub-optimal. Therefore, we propose a novel Multi-Scenario ReID (**MS-ReID**) framework with multiple classification tokens and a scenario-aware identity loss to facilitate effective learning of specific features for each scenario. To achieve better discriminative feature extraction for different scenarios, we improve MS-ReID at both the model structure and optimization levels. Specifically, we propose a Mixture-of-Attribute-Experts (**MoAE**) module, which builds the expert network and assigns different experts to address distinct scenarios, thus enabling the model to alleviate interference between scenarios. Additionally, we define the attribute layer as the basic cell shared among experts with similar scenario attributes, e.g., DT-related attribute layers are shared among DT-LT and DT-ST experts. With this, the model can benefit from multiple interrelated scenarios. And we propose a Hierarchical Dynamic Weighting (**HDW**) strategy, that tackles the AT-ReID training from the multi-task learning view. It establishes all scenarios into several tasks and balances the training for different tasks with a loss weighting scheme. This method considers multiple relevant tasks when computing weights, implicitly modeling the relationships between tasks and leading to better optimization of the overall multi-scenario learning framework.

Our main contributions can be summarized as follows:

- We investigate a new task called AT-ReID, which aims at enabling retrieval at any time moment and across different time intervals. We contribute for the first time a large-scale dataset named AT-USTC to support the study of AT-ReID. Compared to existing datasets, AT-USTC stands out for its long data collection period and the inclusion of both RGB and IR camera footages, meeting the requirement of AT-ReID. Importantly, our data collection has obtained the consent of each volunteer.
- We propose a Uni-AT model to effectively handle all scenarios of AT-ReID. In Uni-AT, three components, including a new multi-scenario ReID framework, a Mixture-of-Attribute-Experts module, and a Hierarchical Dynamic Weighting training strategy are proposed to tackle the new challenges of multi-scenario retrieval in AT ReID tasks. Extensive experiments show that our model leads to satisfactory results and exhibits excellent generalization to all scenarios.

2 Related Work

Person Re-Identification. Traditional ReID (Tr-ReID) aims to achieve short-term pedestrian retrieval in the RGB modality. The corresponding datasets [Zheng *et al.*, 2015; Li *et al.*, 2014; Wei *et al.*, 2018] focused on providing more identities as well as more camera variations. Some Tr-ReID methods designed stronger backbone networks [Luo *et al.*, 2019; Ye *et al.*, 2021; He *et al.*, 2021] or effective ReID losses [Sun *et al.*, 2020], while others [Sun *et al.*, 2018] explored the use of part-level features to obtain discriminative representations of pedestrians.

Visible-infrared cross-modality ReID (CM-ReID) aims to achieve short-term pedestrian retrieval between the RGB and the infrared (IR) modalities. The corresponding datasets [Wu *et al.*, 2017; Nguyen *et al.*, 2017; Zhang and Wang, 2023] focused on providing more RGB and IR cameras. Some CM-ReID methods [Feng *et al.*, 2023; Yang *et al.*, 2023a] aimed to project features from different modalities into the same feature space, while others [Li *et al.*, 2022; Fang *et al.*, 2023] aimed to learn cross-modality relationships.

Cloth-changing ReID (CC-ReID) aims to achieve long-term pedestrian retrieval in the RGB modality. The corresponding datasets [Yang *et al.*, 2019; Qian *et al.*, 2020; Xu and Zhu, 2023] focused on providing clothing variations for each person. Some CC-ReID methods [Chen *et al.*, 2021; Liu *et al.*, 2023] introduced additional data such as contour, key points, human parsing, and 3D shape for model training, while others [Gu *et al.*, 2022; Li *et al.*, 2023] utilized RGB images only to learn robust clothing-irrelevant feature.

The aforementioned tasks and datasets can only cover a portion of the AT-ReID scenarios. In addition, some unified methods [Chen *et al.*, 2023; Tang *et al.*, 2023; Li *et al.*, 2024] focus on multiple tasks, such as text/sketch-to-RGB ReID,

[1]https://github.com/kw66/AT-ReID.

clothes template-based ReID, and occlusion ReID, as well as human-centric tasks like human parsing and pose estimation. Our research is distinct from previous methods as it is the first to focus on the availability of ReID at any time and proposes a relevant dataset and method to bridge the gap between existing research and AT-ReID.

Multi-Task Learning. Multi-task learning (MTL) refers to building a model that can handle multiple distinct tasks.By sharing parameters between tasks, MTL methods achieve efficient memory and data utilization and expect to derive benefits from multiple related tasks. In AT-ReID, various input modalities and learning objectives are present in different scenarios. Retrieval in each scenario can be considered an individual ReID task, and it is promising that employing MTL methods can improve the overall efficacy of the model across all scenarios.

Some MTL methods focused on network architecture [Ma *et al.*, 2019; Wang *et al.*, 2023] to achieve more effective parameter sharing. Recently, some effective approaches [Ma *et al.*, 2018; Tang *et al.*, 2020; Bao *et al.*, 2022] are to utilize the Mixture-of-Experts (MoE) [Jacobs *et al.*, 1991] model that employs multiple expert sub-networks to tackle multi-task learning. Compared to these MoE methods, our MoAE constructs scenario experts in a more flexible sharing manner, making the model benefit from multiple interrelated scenarios. Other methods focused on MTL optimization, such as manipulating gradient [Liu *et al.*, 2021] and adjusting the loss weight by task difficulty, training speed, and priority [Guo *et al.*, 2018]. Our HDW method groups tasks based on their attributes and applies hierarchical dynamic weighting to the loss of each task, achieving a more effective task balance.

3 AT-USTC Dataset

Dataset Description. AT-USTC is the first AT-ReID benchmark that includes 135,465 (69,791 RGB and 65,674 IR) images of 270 identities and 710 sets of different clothing captured by 16 cameras. We deployed 8 RGB and 8 IR cameras across 16 non-overlapping locations, comprising 5 indoor and 11 outdoor scenes. We filmed videos on 13 dates and 26 different time periods during spring, summer, and winter, with temperatures ranging from -3°C to 33°C to cover a wider range of clothing types. Each individual in our training set has 2-14 outfits with an average of 3.6, which facilitates retrieval in long-term scenarios. Due to the variations in both modality and clothing in AT-USTC, the process of capturing and annotating the data is more time-consuming compared to other datasets. We made considerable effort to provide annotations, including labels for person, camera, and clothing.

Dataset Advantages. As shown in Fig. 2 (a), our AT-USTC captured images in various scenarios, including daytime, nighttime, indoor, outdoor, and cross-season intervals, providing comprehensive variations of modality, clothing, camera, and scene. In contrast, the three Tr-ReID datasets, Market1501, CUHK03, and MSMT17 do not include clothing changes and IR cameras; the three CM-ReID datasets, RegDB, SYSU-MM01, and LLCM do not consider clothing changes; the three CC-ReID datasets, PRCC, LTCC, and DeepChange do not include IR cameras.

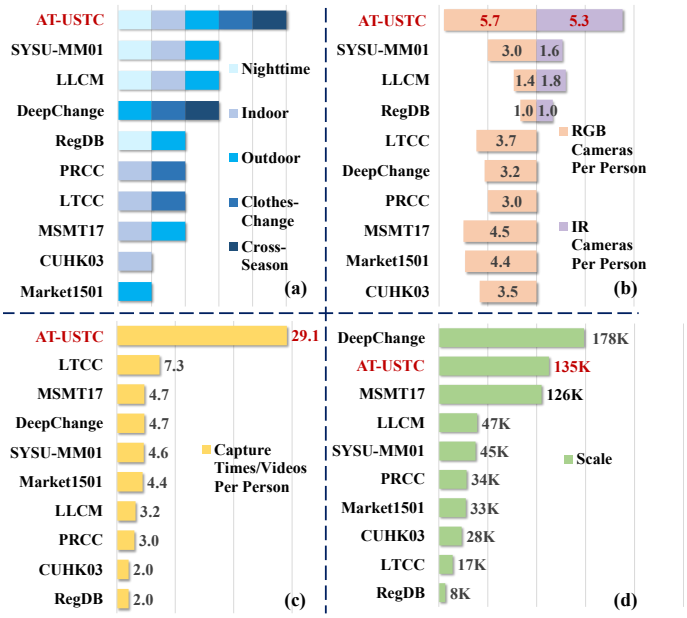

Figure 2: Statistics of our AT-USTC and several popular ReID datasets. Synthetic datasets, datasets from movies [Shu *et al.*, 2021], and the internet [Yıldız and Kasım, 2024] are not within the scope of comparison.

As shown in Fig. 2 (b), we provide rich camera variations for each person during day and night, facilitating cross-camera and cross-modality retrieval. On average, each person of AT-USTC appears in 5.7 RGB and 5.3 IR cameras, totaling 11 cameras, which is significantly higher than other datasets.

As illustrated in Fig. 2 (c) and Fig. 3, each person in AT-USTC has been photographed 29.1 times on average, which is 9 / 6 / 6 times more than large-scale CM-ReID / CC-ReID / Tr-ReID dataset, LLCM / DeepChange / MSMT17. Therefore, this results in significantly higher intra-identity diversity and visual variations of our AT-USTC dataset. Compared with existing datasets, our AT-USTC exhibits day and night photography, diverse cameras, and captures across multiple seasons, enabling it to encompass all scenarios of AT-ReID.

As shown in Fig. 2 (d), the scale of our AT-USTC exceeds that of most other datasets, owing to the higher intra-identity diversity exhibited by each individual in the dataset.

Data Split. We have a fixed split of the dataset into training and testing sets. The training set consists of 135 people with 109,183 images, and the testing set consists of another 135 people with 26,510 images. We partitioned 20% images from the training set for validation purposes. Existing datasets primarily evaluate a single scenario, while we construct separate gallery sets and query sets for all six scenarios covered by AT-ReID to explore anytime retrieval.

4 Method

Overview. Anytime ReID (AT-ReID) aims to perform retrieval in multiple scenarios, including daytime short-term (DT-ST), daytime long-term (DT-LT), all-day short-term (AD-ST), all-day long-term (AD-LT), nighttime short-term (NT-ST), and nighttime long-term (NT-LT) scenarios.

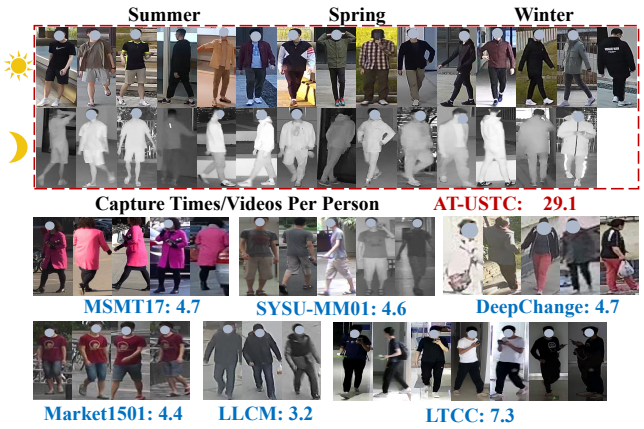

Figure 3: Each person in AT-USTC has been photographed 29.1 times on average, resulting in higher intra-identity diversity. Images of each dataset all belong to the same individual.

The pipeline of our proposed Unified AT-ReID model (Uni-AT) is shown in Fig. 4. The image is fed into a Multi-Scenario ReID (MS-ReID) framework to extract several types of scenario features for accurate retrieval in all covered scenarios of AT-ReID. To treat each scenario optimally, we further propose a Mixture-of-Attribute-Experts (MoAE) module to effectively capture scenario-specific clues and mitigate inter-scenario interference. To balance feature learning in different scenarios, we proposed a Hierarchical Dynamic Weighting (HDW) scheme to train the whole model more effectively in an end-to-end way.

4.1 Multi-Scenario ReID

Model Architecture. The proposed Multi-Scenario ReID (MS-ReID) framework is designed to extract multiple scenario-specific features more effectively. We choose Vision Transformer (ViT) [Dosovitskiy *et al.*, 2020] as our backbone. The input image x_i is split into patches and mapped to patch tokens. Then we establish 6 CLS tokens t_i^s to extract image features and assign each one with a corresponding scenario s. These CLS tokens serve as information gatherers, collecting different scenario-specific knowledge from patch tokens by stacked self-attention modules.

Note that the main principle of our MS-ReID is to extract different features for different scenarios separately rather than use a single unified representation for all scenarios because the latter sacrifices specific clues in each scenario. The main concern is based on a prior that scenario-specific information can lead to optimal results under specific cases. For example, color information is suitable for daytime retrieval and clothes cues are discriminative for short-term situations. Moreover, AT-ReID is a scenario-determinable task, as the practical ReID involves retrieving between the query image and the gallery images in the video surveillance, and the shooting timestamps can be easily accessed. When faced with an uncertain scenario, the default is set to all-day/long-term to account for potential modality variations/clothing changes. Thus, our framework is adaptable and capable of offering more precise solutions for determined scenarios.

Scenario-aware identity loss. The common identity loss provides undifferentiated supervision for all scenarios, which can only learn the shared information across all scenarios but cannot capture scenario-specific cues. Therefore, we propose a scenario-aware identity loss $\mathcal{L}_{id}^s$ for our MS-ReID framework to supervise feature learning for each covered scenario, where different scenarios have non-shared classifiers, distinct negative category sets, and different modality filtering mechanisms. $\mathcal{L}_{id}^s$ is derived as follow:

$$\mathcal{L}_{id}^s(t_i^s) = -\log(p_{i,gt}^s), \tag{1}$$

where $p_{i,gt}^s$ is the classification probability for scenario s. $p_{i,gt}^s$ is derived as follow:

$$p_{i,gt}^s = \frac{\exp(o_{i,gt}^s)}{\exp(o_{i,gt}^s) + \sum_{j \in N_s} \exp(o_{i,j}^s)}, \tag{2}$$

where o_i^s is classification logit generated from the scenario CLS token t_i^s and N_s is the negative category set for the scenarios s. We employ distinct N_s for ST-related and LT-related scenarios. Specifically, for ST cases, we treat different clothes as different categories in classification to guide the model to attend to fine-grained discriminative information about clothes. However, we do not expect the model to classify the same person's images with different clothes into different categories because they share consistent semantic body information. To achieve this goal, we set N_s of ST cases as the clothes ID set while each clothes in this set has different owners with the ground-truth clothes. For the LT cases, we follow the traditional ReID setting to define the category space as the set of person IDs. So N_s of LT cases are different ID persons directly. In addition to distinguishing between ST and LT scenarios, scenario-aware identity loss also includes a modality filtering mechanism. For RGB images, we supervise the DT-ST, DT-LT, AD-ST, and AD-LT tokens of their potential scenarios while neglecting the NT-related ones that are not available to RGB images. Similarly, for IR images, we ignore their DT-related tokens.

Our MS-ReID can fit the goal of each scenario separately, facilitating multi-scenario learning on a dataset with modality and clothing variations.

4.2 Mixture-of-Attribute-Experts

In our MS ReID framework, CLS tokens are supervised by distinct identity loss, aiming to extract corresponding scenario-specific features. However, similar to the case of multi-task learning [Meyerson and Miikkulainen, 2017], the parameter-shared ViT feature extraction network may result in potential gradient conflicts. For instance, the all-day scenario focuses solely on modality-shared information while the daytime scenario needs to also consider RGB-specific information, leading to mutual interference in feature learning between scenarios.

To tackle this problem, inspired by Mixture-of-Experts (MoE) methods [Jacobs *et al.*, 1991; Ma *et al.*, 2018], we propose a novel Mixture-of-Attribute-Experts (MoAE) module. As depicted in Fig. 4, the MoAE module is added parallel with the feed-forward network (FFN) in the transformer layers. In each transformer layer, patch tokens are fed into the

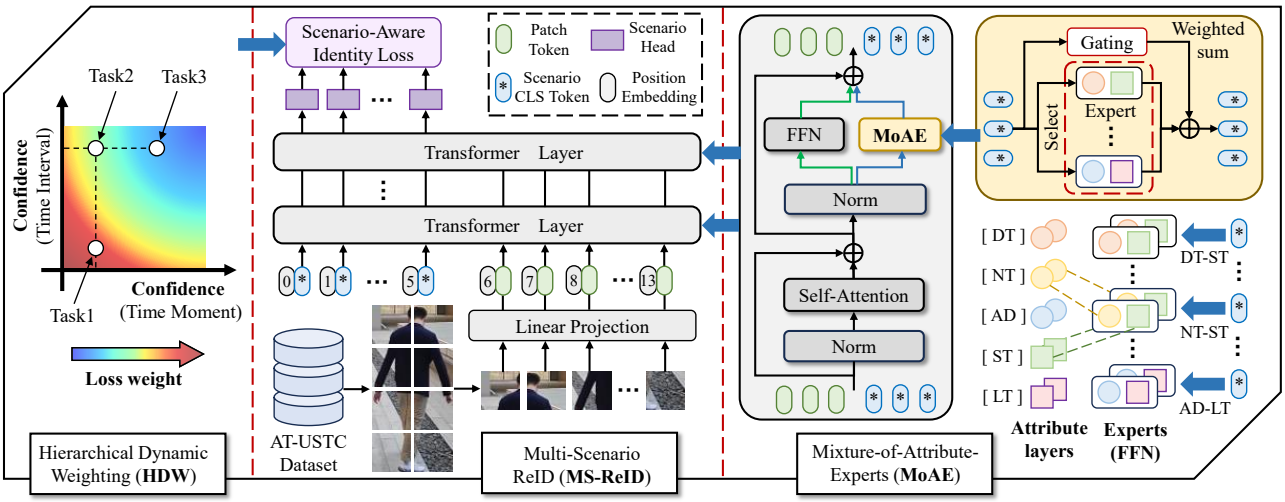

Figure 4: The pipeline of our Uni-AT. The DT, NT, AD, ST, and LT denote daytime, nighttime, all-day, short-term, and long-term cases, respectively. With a Multi-Scenario ReID framework, a Mixture-of-Attribute-Experts module, and a Hierarchical Dynamic Weighting scheme, Uni-AT enhances the learning of diverse scenario-specific features and improves model generalization.

original FFN, while the CLS tokens are fed into the MoAE module. We establish MoAE with $6n$ experts and assign different experts to address distinct scenarios. For each scenario s, there are n specific experts $\{E_j^s\}_{j=1}^n$, where E_j^s represents a single-layer FFN comprising two linear layers and a gelu [Hendrycks and Gimpel, 2016] activation function. Given an input CLS token $t_i^s \in R^d$ corresponding to the scenario s, the MoAE selects and combines experts through a gating network, producing the output $y \in R^d$. More precisely, y is the weighted sum of the outputs from the n experts:

$$y = \sum_{j=1}^{n} G^s(t_i^s)_j E_j^s(t_i^s), \tag{3}$$

where $G^s(t_i^s) \in R^n$ is the weight of n experts, calculated by a gating network:

$$G^s(t_i^s) = top_k(\text{softmax}(W_g^s \cdot t_i^s)), \tag{4}$$

where, $W_g^s \in R^{n \times d}$ is the trainable weights for scenario s in gate decision, and the $top_k(\cdot)$ operator sets all values to zeros except the top-k largest values. Except for the selected k experts, others do not need to be computed for saving computation. With this module, different inputs can utilize experts in distinct ways and capture scenario-specific clues.

Additionally, we note the existence of potential relationships among various scenarios. For example, DT-ST and DT-LT scenarios both have the "DT" attribute, indicating that they require the model to extract RGB-specific features. To introduce this knowledge and establish relationships across scenarios, we construct five types of attribute layers as the basic cells shared among experts with similar scenario attributes. Among these, "DT", "AD", and "NT" attribute layers belong to the category of Time-Moment (TM), while "ST", and "LT" attribute layers belong to the category of Time-Interval (TI). In practice, attribute layers are linear layers and we derive the experts by combining one TM attribute layer and one TI attribute layer as follows:

$$E^s(\cdot) = a_1(gelu(a_2(\cdot))), \tag{5}$$

where a_1 and a_2 are the attribute layers associated with scenario s. In summary, our attribute layers are shared across different scenarios, which can serve as prior knowledge for the relationships between scenarios. Compared to other MoE methods [Ma *et al.*, 2018; Tang *et al.*, 2020; Bao *et al.*, 2022], our MoAE is more flexible and effective for expert construction, which lets the model benefit from multiple interrelated scenarios.

4.3 Hierarchical Dynamic Weighting

In our MS-ReID framework, we utilize scenario-aware identity loss to supervise each scenario, which can be considered as multi-task learning. Through joint learning across multiple tasks, we aim to increase the model's generalization ability in various scenarios. However, different tasks have different levels of difficulty and learning curves. Simply summing all losses with fixed weights to optimize the overall framework may not provide adequate training for some tasks, leading to limited robustness.

To achieve a balanced optimization across all tasks, we propose a Hierarchical Dynamic Weighting (HDW) scheme. The idea of HDW is that, when some tasks retrieve corresponding images with low predicted confidence, they should contribute more to the final loss, and vice versa. The HDW can be exported as follows:

$$\mathcal{L}_{total} = \sum_{s \in S} w^s \cdot \mathcal{L}_{id}^s, \tag{6}$$

where w^s is the weight of the scenario s. Additionally, we observe that the learning of different tasks is interrelated rather than independent. For instance, the retrieval task in DT-ST scenarios requires RGB-specific features and clothing features. The former can be learned from two DT-related tasks, while the latter can be learned from three ST-related tasks. Therefore, the weight adjustment for these tasks should also adhere to this principle. To achieve this, w^s is calculated by two terms and derived as follows:

$$w^s = w_{tm}^s \cdot w_{ti}^s. \tag{7}$$

Inspired by the Focal Loss [Lin *et al.*, 2017], w^s_{tm} and w^s_{ti} can be computed as follows:

$$w^s_{tm} = (1 - p^s_{tm})^{\frac{\gamma}{2}}, w^s_{ti} = (1 - p^s_{ti})^{\frac{\gamma}{2}}, \tag{8}$$

where γ is the hyper-parameter, p^s_{tm} and p^s_{ti} corresponding time moment and time interval confidences for scenario s, respectively. For example, if s is the DT-LT scenario, p^s_{tm} is the confidence of the DT situation while p^s_{ti} is the LT case confidence. We compute the confidence by averaging the ground truth classification probabilities of CLS tokens from scenarios s' in each training batch:

$$\begin{aligned} p^s_{tm} &= mean(exp(-L^{s'}_{id}(t^{s'}_i))), \quad s'_{tm} = s_{tm}, \\ p^s_{ti} &= mean(exp(-L^{s'}_{id}(t^{s'}_i))), \quad s'_{ti} = s_{ti}, \end{aligned} \tag{9}$$

where scenarios s' carry the same time moment/time interval attribute as target scenario s.

Different from traditional multi-task learning strategies [Guo *et al.*, 2018] assuming each task is independent of the other, our joint weighting method considers multiple relevant tasks when computing weights, implicitly modeling the relationships between tasks and leading to better optimization of the overall multi-task learning framework.

5 Experiments

Datasets. We primarily conducted experiments on the proposed AT-USTC dataset due to the lack of support for multi-scenario training and evaluation of AT-ReID in existing datasets. To further evaluate our AT-USTC dataset and Uni-AT method, we utilized several popular ReID datasets for cross-domain generalization experiments.

Evaluation Protocols. The Rank-k matching accuracy and mean average precision (mAP) are adopted as evaluation metrics. For AT-USTC, we conducted separate tests in six different scenarios, including DT-ST, DT-LT, NT-ST, NT-LT, AD-ST, and AD-LT. The average performance of six scenarios is referred to as **Any-Time**, which evaluates the model's ability to retrieve at any given time. For other datasets, we adhered to the evaluation settings of their original papers.

Implementation Details. We use a ViT-Base model with patch size 16 and step size 16 as our backbone. Following existing work [Luo *et al.*, 2019], we introduce the BNNeck before the classifier. All person images are resized to 256×128 and are augmented with random horizontal flipping, padding, random cropping, and random erasing [Zhong *et al.*, 2020] in training. The batch size is set to 64 with 8 identities. The whole model is trained for 120 epochs (24K iterations) with the SGD optimizer. The learning rate is initialized as 0.008 with the warm-up scheme and cosine learning rate decay. The hyper-parameter k and γ are both set to 1.

5.1 Generalization Evaluation of AT-USTC

One of our main contributions is collecting the AT-USTC dataset, which exhibits higher intra-identity diversity compared to existing ReID datasets. To evaluate the quality of our dataset, we conducted domain generalization experiments, comparing it with three large-scale datasets, including MSMT17, DeepChange, and LLCM. For a fair comparison at the dataset level, we trained the same ResNet50 model for all datasets instead of our proposed model. As shown in Tab. 1, the model trained on AT-USTC achieved the best cross-dataset performance, surpassing the model trained on MSMT17 by an average of 9.6% in Rank-1 accuracy. This shows that AT-USTC has excellent scalability and can effectively support research on the AT-ReID task.

Train \ Test	Market 1501	CUHK 03	SYSU-MM01	PRCC	LTCC	AVG
MSMT17	52.1	10.6	3.2	28.2	17.6	22.3
DeepChange	50.6	4.1	2.4	30.1	12.0	19.8
LLCM	39.3	3.7	6.5	26.6	10.2	17.3
AT-USTC	**52.7**	**19.1**	**26.2**	**38.1**	**23.2**	**31.9**

Table 1: Cross-domain generalization experiments in different datasets. Rank-1 accuracy (%) is reported.

Method	Experts	Params	Time	Rank-1	mAP
MS-ReID	-	1.0×	1.0×	50.1	36.7
+MMoE [Ma *et al.*, 2018]	6	4.9×	1.2×	51.1	37.1
	12	8.8×	1.5×	51.8	37.8
+PLE [Tang *et al.*, 2020]	7	5.6×	1.2×	51.6	37.9
	12	8.8×	1.3×	52.0	38.3
+VLMo [Bao *et al.*, 2022]	3	3.0×	1.1×	51.9	37.9
+MoAE (ours)	6	2.6×	1.1×	52.5	38.4
	12	4.2×	1.2×	**53.2**	**38.7**

Table 2: Comparison with other MoE methods on AT-USTC.

5.2 Comparison with MoE Methods

As shown in Tab. 2, we compare our MoAE module with other MoE methods under Any-Time testing on the AT-USTC dataset. For a fair comparison, all MoE modules are added to our MS-ReID framework in the same manner. The difference among these methods lies in the sharing of experts. MMoE constructs experts shared across all scenarios, while PLE explicitly separates the experts into scenario-specific ones and those shared across all scenarios. VLMo was originally designed to process visual and language modalities. Here, we apply its principles to handle the RGB and IR modalities in the ReID task, establishing three experts for the RGB, IR, and cross-modality data.

From the experimental results, it can be summarized that our MoAE consistently outperforms other methods when the number of experts or parameters is comparable. This is because our MoAE introduces scenario priors and enables fine-grained expert sharing, whereas other methods can only coarsely share experts across scenarios.

Furthermore, our training or inference speed is only 1.2 times that of the non-expert model when using 12 experts because we only select the top-k (k=1) experts for computation. Through parameter sharing with our attribute layers, MoAE not only effectively extracts scenario-specific features to enhance performance but also exhibits greater parameter and time efficiency than other MoE methods.

5.3 Ablation Analysis

As shown in Tab. 3, we conduct ablation studies to evaluate each component of our method. In the 1-st row, we establish an MS-ReID baseline using the standard identity loss for

Method			AT-USTC Dataset							Cross-domain Generalization		
L_{id}^{s}	MoAE	HDW	Any-Time R1/mAP	DT-ST R1/mAP	DT-LT R1/mAP	NT-ST R1/mAP	NT-LT R1/mAP	AD-ST R1/mAP	AD-LT R1/mAP	Market R1/mAP	SYSU R1/mAP	LTCC R1/mAP
✗	✗	✗	48.8/34.1	91.8/78.2	27.8/17.9	71.3/42.3	41.3/26.4	35.3/22.6	25.0/17.2	62.7/36.0	19.6/20.1	24.3/13.9
✓	✗	✗	50.1/36.7	94.9/84.2	26.2/17.4	74.3/47.2	40.2/27.3	39.3/26.0	25.9/18.3	65.7/40.5	24.5/24.6	26.5/14.1
✓	✓	✗	53.2/38.7	95.6/85.5	30.0/19.1	77.6/49.4	43.9/28.3	43.7/30.7	27.5/19.1	66.8/42.3	25.5/25.6	30.3/15.1
✓	✗	✓	52.1/38.7	95.8/86.8	28.6/18.3	76.8/49.7	40.5/27.3	44.8/31.0	26.1/19.0	70.6/45.5	27.5/27.7	28.8/15.3
✓	✓	✓	**54.4/40.0**	**96.7/87.5**	**31.4/21.0**	**79.0/50.2**	**43.1/28.7**	**47.2/32.9**	**29.2/19.5**	**71.5/46.1**	**28.4/28.1**	**31.0/16.0**

Table 3: Ablation study about each component. Rank-1 (R1) and mAP accuracy (%) are reported.

Task	Method	Any-Time		DT-ST		DT-LT		NT-ST		NT-LT		AD-ST		AD-LT	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Tr-ReID	PCB [Sun *et al.*, 2018]	46.7	31.4	91.9	74.5	25.4	16.8	68.8	40.5	35.2	18.8	35.7	22.3	23.3	15.6
	TransReID [He *et al.*, 2021]	50.8	35.5	94.1	81.2	28.9	18.9	72.8	42.0	40.7	26.3	39.5	26.0	28.8	18.5
CC-ReID	CAL [Gu *et al.*, 2022]	51.6	36.0	93.3	79.7	30.2	20.5	75.5	45.1	40.0	25.1	43.0	28.2	27.3	17.1
	AIM [Yang *et al.*, 2023b]	51.0	35.4	92.9	78.0	29.4	20.3	74.6	44.4	40.2	24.8	41.5	27.4	27.5	17.6
CM-ReID	AGW [Ye *et al.*, 2021]	48.8	32.3	91.8	72.5	27.9	17.7	73.1	43.1	37.3	21.5	38.2	24.0	24.4	14.7
	CIFT† [Li *et al.*, 2022]	51.8	36.1	92.4	79.3	29.7	20.0	74.0	43.9	41.1	25.6	45.5	30.6	28.3	17.4
	SAAI† [Fang *et al.*, 2023]	52.3	36.8	94.0	82.1	29.3	19.2	74.8	45.0	41.5	25.8	45.9	30.8	28.5	17.6
AT-ReID	baseline	48.8	34.4	91.8	80.0	27.8	17.9	71.3	42.3	41.3	26.4	35.3	22.6	25.0	17.2
	Uni-AT (Ours)	**54.4**	**40.0**	**96.7**	**87.5**	**31.4**	**21.0**	**79.0**	**50.2**	**43.1**	**28.7**	**47.2**	**32.9**	**29.2**	**19.5**

Table 4: Comparison with the state-of-the-arts on the AT-USTC dataset. For these methods, we searched for hyper-parameter values suitable for AT-USTC to achieve better performance. † implies that no sample relationship from the gallery is used.

six CLS tokens, which provides undifferentiated supervision for all scenarios, failing to take advantage of our MS-ReID framework. In the 2-nd row, we use scenario-aware identity loss L_{id}^{s} to facilitate the effective knowledge acquisition from each scenario, leading to improvements of 1.3% and 2.6% in Rank-1 and mAP accuracy on Any-Time testing. However, this approach tacks each scenario independently without considering their relationships, limiting efficient collaborative optimization across multiple scenarios.

To address this issue, we further introduce the MoAE and HDW components displayed in the 3-rd and 4-th rows. The integration of MoAE results in notable improvements of 3.1% and 2.0% in Rank-1 and mAP accuracy. The MoAE module introduces an expert mechanism to guide the model in capturing discriminative scenario-specific cues, effectively improving feature learning across relevant scenarios. Furthermore, the incorporation of the HDW scheme improves Rank-1 and mAP accuracy by 2.0% and 2.0%. This scheme facilitates balanced learning of specific features across various scenarios in an end-to-end manner. Ultimately, as shown in the 5-th row, combined of all components leads to remarkable performance enhancements of 5.6% and 5.9% in Rank-1 and mAP accuracy on Any-Time testing, and 8.8% / 8.8% / 6.7% and 9.9% / 8.0% / 2.1% improvements in Rank-1 and mAP accuracy for cross-dataset testing on the Market1501 / SYSU-MM01 / LTCC dataset. This demonstrates the complementary of the proposed L_{id}^{s}, MoAE, and HDW, making the entire MS-ReID framework more effective.

5.4 Comparison with ReID Methods

As shown in Tab. 4, we compare our method with the SOTAs of Tr-ReID, CC-ReID, and CM-ReID tasks. Our method achieves the highest performance, surpassing the second-best method SAAI by 5.4%, 1.8%, 5.2%, 2.9%, 2.1%, and 1.9% in mAP accuracy on the six AT-ReID scenarios. This demonstrates that existing methods designed for a target scenario are inadequate for addressing the challenges of the multi-scenario AT-ReID. These methods learn a unified representation for all scenarios and assign all images of a person to a single category regardless of modality or clothing. Consequently, they focus solely on shared features across six scenarios while neglecting specific cues, resulting in poor generalization in non-target scenarios and diminished retrieval performance in target scenarios. Our Uni-AT approach employs a novel multi-scenario learning framework to capture specific features of different scenarios and facilitate mutual enhancement among scenarios, which exhibits notable advantages compared to the baseline and other ReID methods.

6 Conclusion

In this paper, we investigate a new ReID task, Anytime ReID (AT-ReID), and introduce the first corresponding dataset named AT-USTC to support its research. Compared to existing datasets, AT-USTC is characterized by its extensive intra-identity diversity in clothing, modality, and camera, effectively fulfilling the crucial requirements of AT-ReID. We analyze this new challenge task and propose a Unified AT-ReID model to handle the multi-scenario AT-ReID through a single model. By improving model architecture and optimization, we achieve accurate retrieval results across multiple scenarios. The proposed methods and datasets may inspire the following ReID methods towards real application.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62272430).

Contribution Statement

Xulin Li and Yan Lu made equal contributions to this work.

References

[Bao *et al.*, 2022] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[Chen *et al.*, 2021] Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, and Wei-Shi Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8146–8155, 2021.

[Chen *et al.*, 2023] Cuiqun Chen, Mang Ye, and Ding Jiang. Towards modality-agnostic person re-identification with descriptive query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15128–15137, 2023.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fang *et al.*, 2023] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11270–11279, 2023.

[Feng *et al.*, 2023] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22752–22761, 2023.

[Gu *et al.*, 2022] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1069, 2022.

[Guo *et al.*, 2018] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European Conference on Computer Vision*, pages 270–287, 2018.

[He *et al.*, 2021] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[Jacobs *et al.*, 1991] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[Li *et al.*, 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[Li *et al.*, 2022] Xulin Li, Yan Lu, Bin Liu, Yating Liu, Guojun Yin, Qi Chu, Jinyang Huang, Feng Zhu, Rui Zhao, and Nenghai Yu. Counterfactual intervention feature transfer for visible-infrared person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 381–398. Springer, 2022.

[Li *et al.*, 2023] Xulin Li, Yan Lu, Bin Liu, Yuenan Hou, Yating Liu, Qi Chu, Wanli Ouyang, and Nenghai Yu. Clothes-invariant feature learning by causal intervention for clothes-changing person re-identification. *arXiv preprint arXiv:2305.06145*, 2023.

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

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[Liu *et al.*, 2021] Liyang Liu, Yi Li, Zhanghui Kuang, J Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. iclr, 2021.

[Liu *et al.*, 2023] Fangyi Liu, Mang Ye, and Bo Du. Dual level adaptive weighting for cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 2023.

[Lu *et al.*, 2020] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020.

[Luo *et al.*, 2019] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[Ma *et al.*, 2018] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-

of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939, 2018.

[Ma *et al.*, 2019] Jiaqi Ma, Zhe Zhao, Jilin Chen, Ang Li, Lichan Hong, and Ed H Chi. Snr: Sub-network routing for flexible parameter sharing in multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 216–223, 2019.

[Meyerson and Miikkulainen, 2017] Elliot Meyerson and Risto Miikkulainen. Beyond shared hierarchies: Deep multitask learning through soft layer ordering. *arXiv preprint arXiv:1711.00108*, 2017.

[Nguyen *et al.*, 2017] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[Qian *et al.*, 2020] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[Shu *et al.*, 2021] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4390–4403, 2021.

[Sun *et al.*, 2018] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*, pages 480–496, 2018.

[Sun *et al.*, 2020] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020.

[Tang *et al.*, 2020] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 269–278, 2020.

[Tang *et al.*, 2023] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21970–21982, 2023.

[Wang *et al.*, 2023] Zhen Wang, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Huan Sun, and Yoon Kim. Multitask prompt tuning enables parameter-efficient transfer learning. *arXiv preprint arXiv:2303.02861*, 2023.

[Wei *et al.*, 2018] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[Wu *et al.*, 2017] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5380–5389, 2017.

[Xu and Zhu, 2023] Peng Xu and Xiatian Zhu. Deepchange: A long-term person re-identification benchmark with clothes change. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11196–11205, 2023.

[Yang *et al.*, 2019] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2029–2046, 2019.

[Yang *et al.*, 2023a] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11069–11079, 2023.

[Yang *et al.*, 2023b] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1472–1481, 2023.

[Ye *et al.*, 2021] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Yıldız and Kasım, 2024] Serdar Yıldız and Ahmet Nezih Kasım. Entire-id: An extensive and diverse dataset for person re-identification. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–5. IEEE, 2024.

[Zhang and Wang, 2023] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[Zhong *et al.*, 2020] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.