A Multi-Background Normalization and Dynamic Meta Feature Mining Approach for Person Re-Identification

Xiaohao Xie

Abstract

Person re-identification (ReID) aims to retrieve pedestrians across cameras, facing challenges from differences in perspective, background, and lighting, which introduce noise and hinder key feature extraction. Existing methods, often relying on normalization or generative data augmentation, suffer from limitations such as neglecting camera label information or the unreliability of two-stage learning. To address this, we propose a one-stage architecture, M-MBNNet, consisting of MBN (Multi Background Norm) and MetaRep (Meta-Representation for Adaptive Metric) modules. MBN uses a camera-wise Assignment Gate and Multi-aggregation Norm to align and normalize backgrounds, reducing interference and enhancing person-relevant feature robustness. MetaRep bridges representation and metric learning, leveraging mutual information (quality measures) to dynamically adjust asymmetric metrics for consistent multi-task convergence. It also incorporates curriculum learning to dynamically emphasize either inter-class separability or intra-class compactness. M-MBNNet offers a systematic approach to extracting key pedestrian features and resolving cross-camera differences through active alignment and adaptive optimization. We achieve strong results on two baselines—one mainly for representation and one for metric learning—demonstrating the method's scalability.

1. Introduction

Person re-identification [15] (ReID) aims to retrieve pedestrians across cameras. Cross-camera scenarios introduce significant variations in perspective [11], background , lighting [4], and other factors, making it a pressing challenge for ReID and similar vision tasks to optimize their methods. Beyond addressing camera-related differences, the process of feature learning involves additional challenges: how to align multi-task [3] convergence objectives between representation learning and metric learning, how to emphasize features highly relevant to pedestrian identities while ignoring low-confidence features, and how to enhance the discriminative power [16] of the feature space.

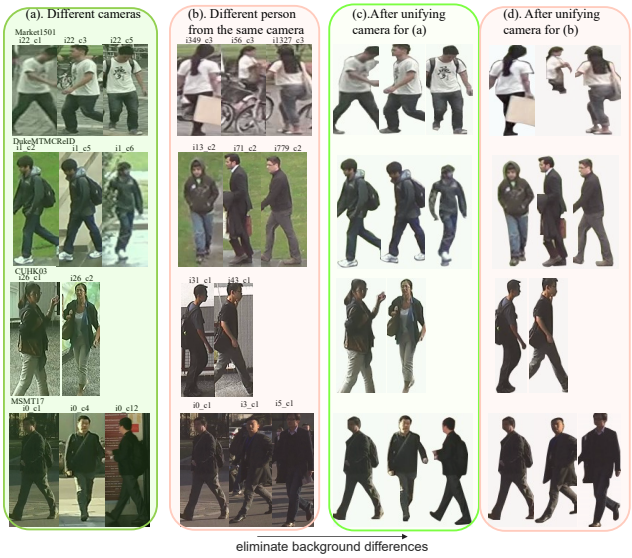

Figure 1. a. Reflect the background variations introduced by different cameras. b. Highlight the strong semantic consistency of images of the same pedestrian captured by the same camera. c & d. Demonstrate how we align images to approximate a Gaussian white-noise background by eliminating camera-specific variations.

1.1. challenges in cross-camera scenarios

Some studies [13] have acknowledged the influence of non-pedestrian factors (e.g., backgrounds captured by different cameras), while others employ generative methods [9] to create new images of pedestrian instances in varied backgrounds, potentially introducing adverse effects and additional noise. Existing cross-camera ReID methods often neglect camera label information during training, merely exposing networks to pedestrian instances from diverse cameras with varying backgrounds. While this implicitly pressures models to focus on pedestrian features, such approaches lack systematic mechanisms to handle background interference – a critical challenge since camera-specific perspectives, lighting and backgrounds inherently inject noise into feature representations. We argue discriminative pedestrian-specific learning requires dual-phase optimization. First, networks should actively acquire camera background recognition capabilities through explicit clas-

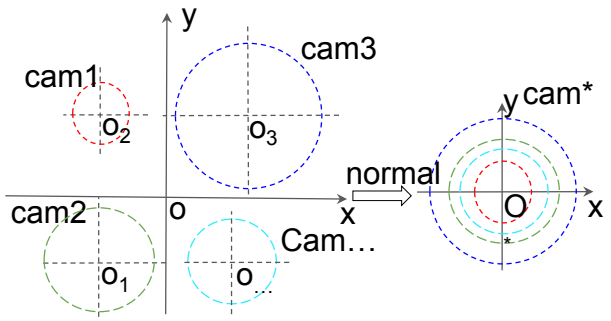

Figure 2. Simulation of MBN. It realizes the camera discrepancyand aligns the cameras.

sification layers or implicit metric learning (e.g., camera-aware center loss [21] or contrastive loss), leveraging the limited yet distinct camera-specific background patterns. Second, the framework need facilitate background-invariant projection by leveraging learned camera characteristics to drive adversarial filtering or style normalization, actively suppressing background perturbations. This methodology diverges from empirical background removal approaches, systematically addressing cross-camera variance through targeted camera-aware representation learning.

1.2. Challenges in Multi-Task Learning Convergence and Feature Discrimination

In cross-camera scenarios, the solution [27] aims to overcome domain biases from different data sources and unify them into a single domain. However, this process may inadvertently erode or weaken certain key discriminative features for distinguishing pedestrians. In representation learning, relative entropy [8] is used to explore the distinctions in data distribution, while metric learning employs distance functions to identify separability in the metric space. Although both methods aim to extract data discriminability, the significant convergence inconsistency between these two learning tasks can weaken the ability of ranking tasks to differentiate key pedestrian features. Therefore, during feature learning, it is crucial to establish a bridge architecture that unifies representation and metric learning, emphasizing the extraction of key discriminative features related to pedestrian identities. **Contributions**. In this paper, we propose M-MBNNet, a novel system architecture designed to extract critical pedestrian-specific features. The core of this architecture lies in two modules: MBN and MetaRep.

1. MBN addresses camera-induced background differences via two sub-modules: the Camera-wise Assignment Gate, enabling asymmetric background learning, and the Multi-Aggregation Norm Block, which aligns backgrounds while preserving pedestrian features.
2. MetaRep employs representation learning to generate meta-representations with mutual information, guiding metric learning by dynamically adjusting margins. It follows a curriculum learning approach, first focusing on inter-class granularity, then intra-cluster fine granularity, enhancing consistency between metric and representation learning. This system robustly extracts critical pedestrian-specific features.

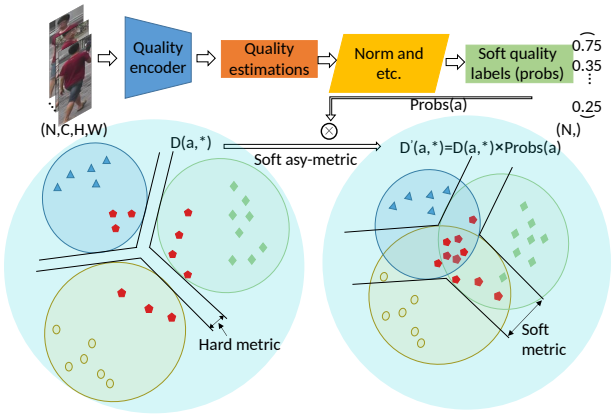

Figure 3. Illustration of MetaRep. Quality evaluation dynamically adjusts adaptive granularity margins to enhance pedestrian feature discriminability. A softened meta-representation serves as the evaluation factor, and an asymmetric metric bridges representation and metric for multi-task convergence.

2. Related Work

2.1. Background/Camera-aware ReID

Several ReID methods have been proposed to address cross-camera differences. In DAPRH [14], GAN-based methods are used to generate new images of pedestrian instances in various backgrounds as input, effectively balancing the distribution of pedestrian data across different background styles through data augmentation. Intra-Camera [26] uses independently annotated labels within single camera views and proposes a multi-task multi-label learning approach to discover identity correspondences across cameras. MgCA-Model [18] employs binary segmentation masks to create RGB-Mask pairs and designs a mask-guided contrastive attention model (MGCM) to separately learn features from body and background regions, ensuring feature embeddings focus on pedestrian features while reducing background interference. In msi-ReID [2], a spatial alignment module (SPM) is introduced to help the network focus on identity-related features during training, ignoring non-identity factors like lighting and background. CBN [27], a camera-based normalization approach leverages camera label information, significantly improving performance. However, it has limitations, such as the inability to actively learn background information and insufficient refinement of normalization layers, which may lead to performance degradation. Additionally, generative methods [14] expand pedestrian image data in specific camera backgrounds to balance background frequency. However, the two-stage training process is complex, and the reliability of the method depends on the accuracy of the generated data.

2.2. Discriminative feature for ReID

Extracting discriminative features is a core objective in person re-identification (ReID) and related tasks. Early methods relied on handcrafted features (e.g., color histograms, texture), but these struggled with environmental and viewpoint variations. DeepReID [5] pioneered CNN-based end-to-end feature learning, significantly boosting ReID performance. Recent advances focus on enhancing feature robustness and discriminability. Metric learning methods like Triplet Loss [17] and Contrastive Loss [1] optimize distance relationships between feature vectors, while rank-in-rank loss [23] addresses class imbalance. Viewpoint-aware contrastive loss [6] learns viewpoint-invariant features, and background interference is mitigated by Background Consistency Constraint (BCC) and Object-Centric Feature Refinement (OCFR) losses [24]. TSNT [8] develop noise-robust mechanism for reliability discrimination. Attention-based methods (e.g., BAM [12], CBAM [22]) and self-attention networks (e.g., Transformers) further improve feature representation by focusing on discriminative image regions. Techniques like PCB (Part-based Convolutional Baseline) [20] and MSINet [2] refine attention consistency across images, significantly boosting ReID accuracy.

2.3. Multitask Consistent Convergence for ReID

Multitask Consistent Convergence refers to ensuring that multiple tasks (e.g., representation and metric learning) co-optimize within a unified framework, achieving stable and aligned convergence to enhance overall model performance. PMT-Net [25] proposes a progressive multi-task network architecture to enhance recognition capabilities by gradually integrating complementary tasks. Huang et al. [3] designed a multi-task learning framework that constructs separate branches for distinct recognition tasks, effectively capturing discriminative body information in cross-modality ReID scenarios. More fundamentally, existing ReID methods often combine representation learning and metric learning losses, but their inconsistent convergence boundaries can lead to suboptimal solutions. To address this, we propose MetaRep, which bridges representation and metric learning with adaptive soft margins [7], replacing fixed margins. MetaRep leverages metadata enriched with mutual information from representation learning to guide metric learning. It incorporates curriculum learning [19], progressing dynamically from easy to hard tasks, and adapts step sizes for both representation and metric learning. During training, MetaRep prioritizes inter-class separability in early stages and shifts to intra-class compactness and fine-grained semantic relationships as the model converges. This approach achieves refined, discriminative feature representations tailored for ranking tasks.

3. Method

We propose an end-to-end person re-identification (ReID) model composed of two core modules: Multi-Background Normalization (MBN) and Meta-Representation Learning for Adaptive Metrics (Meta-Rep). This model addresses cross-camera background interference and the inconsistency between representation learning and metric learning through proactive background alignment and an integrated feedback mechanism.

3.1. Multi-Background Normalization (MBN)

In cross-camera scenarios, differences in camera viewpoints and backgrounds introduce significant noise. The MBN module mitigates this issue through two key components:

Camera-wise Assignment Gate (CAG). This component employs an asymmetric background learning strategy by aligning each sample's features to its corresponding camera's background center. Initially, the distance is computed as:

$$\mathrm{D}(f_i, C_i) = \|f_i - C_i\|_2 \tag{1}$$

To account for the varying importance of samples within the same camera, a dynamic weight ($indicator$) is introduced, modifying the distance to (Asymmetric Center Loss):

$$L_{AsyC} = \mathrm{D}(f_i, C_i) = indicator(x_{c\_i}) \bullet \|f_i - C_i\|_2 \tag{2}$$

The *indicator* is dynamically adjusted based on the model's predicted probability distribution and the corresponding true labels. This mechanism enables the module to evaluate the representativeness of each sample within its respective camera domain. Guided by a curriculum learning strategy, the influence of the indicator is progressively enhanced, ensuring more accurate background alignment.

Here, we adopt the fused meta-representations described in Section 3.2 (refer to details) as the basis for the indicator. The central theoretical rationale is to allow the model to synthesize its current understanding and the true labels to dynamically determine the strength of metric learning toward the centers.

Multi-Aggregation Norm(MN). To further reduce statistical discrepancies among different cameras, the Multi-Aggregation Norm employs independent Batch Normalization (BN) layers for each camera:

$$BNs(x) = \begin{array}{l} \frac{x \cdot I(c=i) - \mathrm{E}[x \cdot I(c=i)]}{\sqrt{\mathrm{Var}[x \cdot I(c=i)] +}} \cdot \gamma + \beta, \\ \text{where} \quad \mathrm{i} = 1, 2, \ldots, \mathrm{n} \end{array} \tag{3}$$

The variable *i* represents the camera index, and the formula illustrates that statistics and learned parameters are calculated based on data corresponding to each specific camera. $I(\bullet)$ represents discrete impulse function, which is 1 when the input condition is true, otherwise it is 0.

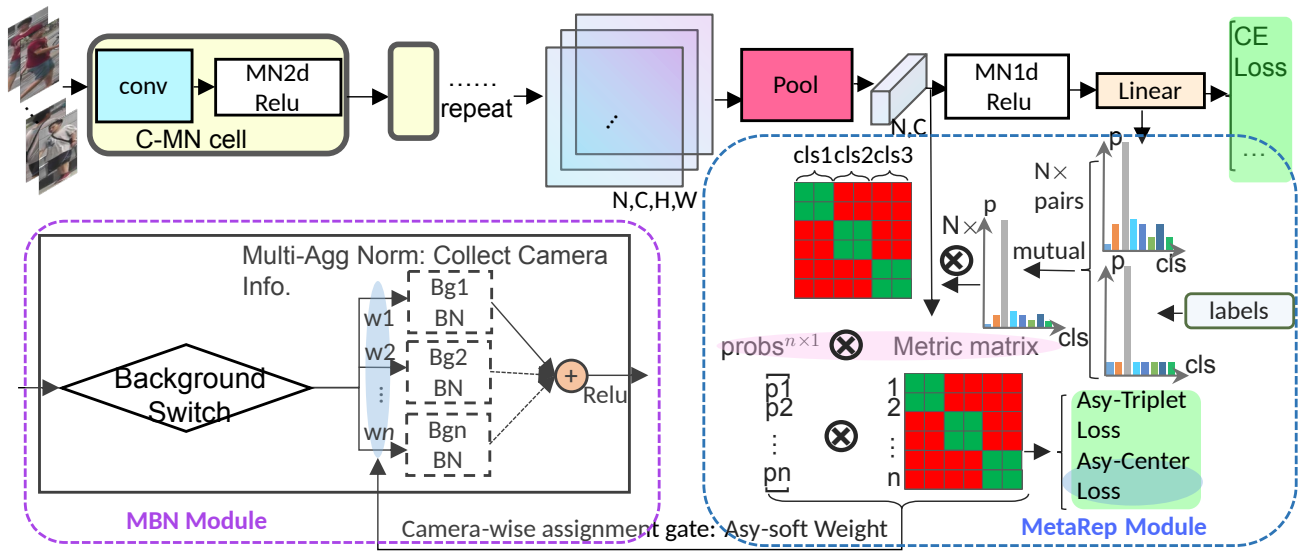

Figure 4. M-MBNNet overview. MBN collect camera-wise information and eliminate the camera Discrepancy. MetaRep leverages soft meta-representations (prob) that incorporate quality measures to perform asymmetric metric learning.

Statistical Considerations Related to MBN. The BN layer approximates overall data statistics using sample estimates. For these estimates to be reliable, the samples must be consistent and numerous. Previous studies have shown that BN performance can degrade when the batch size is too small or when training and inference statistics differ.

Our Multi-Aggregation Norm employs both BN1d and BN2d, as they are well-suited for our CNN backbone, with statistics computed along the channel dimension.

BN1d: For input shaped (N, C) (with N as the batch size and C as feature dimensions):

$$\mu_c = \frac{1}{N}\sum_{n=1}^{N} x_{n,c} \tag{4}$$

$$\sigma_k^2 = \frac{1}{N}\sum_{n=1}^{N} (x_{n,c} - \mu_c)^2 \tag{5}$$

BN2d: For input shaped (N, C, H, W) (where H and W are spatial dimensions):

$$\mu_c = \frac{1}{N \cdot H \cdot W}\sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{n,c,h,w} \tag{6}$$

$$\sigma_c^2 = \frac{1}{N \cdot H \cdot W}\sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W} (x_{n,c,h,w} - \mu_c)^2 \tag{7}$$

BN1d uses N samples per channel, whereas BN2d uses $N \cdot H \cdot W$. Typically, N is limited by memory, and H and W decrease with network depth.

Additionally, BN layers update global statistics with a moving average:

$$\hat{x}_{new} = (1 - momentum) \cdot \hat{x} + momentum \cdot x_t \tag{8}$$

Since overall statistics depend on the dataset scale, we propose dynamically configuring the Multi-Aggregation Norm. If the statistical scale exceeds a threshold $\epsilon \in (3072, 12288)$, the block is used, otherwise discarded.

3.2. MetaRep (Meta-Representation Learning for Adaptive Metrics)

Ranking is one of the core tasks in person re-identification. Its basic form is:

- Input: query image q and candidate set G: $\{g_1, g_2, \ldots, g_n\}$.
- Goal: Sort the query image q and each candidate image according to their similarity, so that the images corresponding to the same person are ranked as high as possible.

The currently popular deep person re-identification model combines representation learning with metric learning. Representation learning extracts robust pedestrian features, and metric learning optimizes similarity metrics to improve ranking performance. Robust pedestrian features promote stable input for metric learning, and reasonable metrics in turn promote the expression of pedestrian features in feature space. The usual combination method is to embed features into a metric space while training the classifier using representation learning methods, and establish a multi-task loss function, such as the loss function of the commonly used ReID strong baseline [10]:

$$L = \lambda_1 \cdot L_{classification} + \lambda_2 \cdot L_{metric} \tag{9}$$

This adopts a simple representation and metric loss function superposition, which is simple and effective, but lacks a direct feedback mechanism between the two tasks, and it is

difficult to converge to a high-quality solution for the rank task.

Representation learning uses relative entropy to measure the similarity between the embedding distribution and the target distribution.Deep metric learning methods often employ margin-based losses, such as contrastive loss and triplet loss (see the formula below). The fundamental idea is to pull samples of the same class closer into a cluster while ensuring that samples of different classes are separated by an absolute or relative distance greater than the margin.

$$L = \max(0, d_{ap} - d_{an} + \alpha), \tag{10}$$

d_{ap}: Anchor-positive distance. d_{an}: Anchor-negative distance. a: Margin parameter.

In a training epoch, representation learning locates the distribution of the current input within the global categories, while metric learning, such as Triplet Loss, is often limited by the batch size and restricted to measuring within a local cluster. As a result, many other clusters are unseen during the current epoch's training. Furthermore, metric learning uses a fixed metric pattern from the start to the end of the training phase (with adjustments primarily to the learning rate, while other parameters remain largely fixed), without following a curriculum learning approach that progresses gradually. Additionally, there is no dynamic adjustment of the margin based on individual samples. Moreover, while representation learning and metric learning are highly collaborative tasks in ReID, they still have very different objectives. Finally, ReID is a ranking task that utilizes metric learning features for instance retrieval, typically discarding the classification layer after representation learning. Recently, ReID generally employs a combined training approach with cross-entropy loss and triplet loss. We view each input's probability distribution as a meta-representation (Meta-Rep), and each distance set of metric inputs (such as the triplet (a, p, n) in triplet loss) as a meta-metric. Within a single training round, the meta-representation measures the input distribution from a global perspective, while the meta-metric evaluates the input's cluster membership from a local perspective.

The standard classification loss (Softmax Loss). For a sample x_i, its classification probability is:

$$P(y = j|x_i) = \frac{\exp(\mathbf{W}_j^\top \mathbf{f}_i + b_j)}{\sum_{k=1}^{C} \exp(\mathbf{W}_k^\top \mathbf{f}_i + b_k)} \tag{11}$$

Where: W is the weight vector of class j, f_i is the feature vector of sample i, and b_j is the bias. The cross-entropy loss (i.e. softmax loss) is:

$$L_{Cls} = -\frac{1}{N} \sum_{i=1}^{N} \log P(y = y_i|x_i) \tag{12}$$

The notation $P(y = j|x_i)$ is abbreviated as P_i^j, which represents the model's confidence in the current sample belonging to class j. P_i denotes the entire probability distribution for the current sample during forward propagation. As the model trains and gradually converges, the model's confidence in $j = Ground\ truth$ increases, indicating that the model's representation fit to the training data improves. However, unlike the deterministic label distribution, the feature distributions of samples from the same class exhibit intrinsic variability even under model convergence. This variability reflects the model's similarity assessment between input samples and class prototypes, a critical mechanism for effective forward propagation. This judgment is particularly crucial when a converged model, trained on the training set, is tested on the test set, as it allows the model to flexibly handle ambiguous samples.

To ensure that the meta-representation contains as much accurate information as possible, we fuse the meta-representation with the label. To ensure physically plausible fused distributions, we apply two normalized weights (summing to unity) to balance the meta-representation and label distribution components. The fused meta-representation is denoted as $Pred$. Since we are only concerned with the mutual information of the correct classification for the current sample, the meta-representation is simplified to binary classification. It can be simplified to:

$$\begin{aligned} Pred &= \lambda_1 \cdot P_i^i + \lambda_2 \cdot y_i^i \\ s.t.\ &\lambda_1 + \lambda_2 \in [0, 1] \wedge \lambda_1 + \lambda_2 = 1 \end{aligned} \tag{13}$$

The above equation is no longer a probability distribution, but a single value.

Following the concept of curriculum learning, progressing from easy to difficult tasks, the idea is to dynamically adjust the learning intensity for each meta-metric and also attempt to dynamically adjust the margin. In fact, a straightforward approach is to multiply the margin α by an adjustment factor, and allow factor to dynamically adjust according to the current learning difficulty, thereby directly adjusting α.

$$L = max(0, d_{ap} - d_{an} + \alpha \cdot factor) \tag{14}$$

In the above equation, $f(x_a)$ represents the feature of anchor a. The derivative of the equation is independent of factor, meaning its influence on features during the optimization process is extremely limited. It only affects the optimization process through the threshold at which the triplet loss takes effect. in practice, factor can be designed as a function of the features, *i.e.*:

$$factor = factor\,(f\,(x)) \tag{15}$$

However, in this case, when calculating gradients, the influence of $factor$ on the distance function d is unclear, which

is unfavorable for metric learning. The factorfactorfactor we introduced aims to dynamically adjust the metric for each triplet (see Eq. (16)), but $factor$ is not integrated multiplicatively with d, meaning it does not directly guide the learning of d during backpropagation. In fact, further modification is straightforward: design a multiplicative factor for d. This factor should indicate the learning intensity of the meta-metric, adjust the margin, and adapt to curriculum learning as the training progresses.

$$L = max(0, factor_{ap} \cdot d_{ap} - factor_{an} \cdot d_{an} + \alpha) \quad (16)$$

Although the above equation does not directly multiply α by a $factor$, scaling d is effectively equivalent to inversely adjusting α. Furthermore, since the anchor a is the core of the meta-metric, $factor_{ap}$ and $factor_{an}$ can be simplified into a single $factor_a$. The remaining task is to design a reasonable $factor_a$. Implementing curriculum learning in this context is straightforward, such as using a scheduler to adjust $factor_a$ over epochs, similar to the paradigm of learning rate scheduling. However, we aim for $factor_a$ to dynamically adjust both the learning intensity of the meta-metric and the margin. This can be achieved by directly using the meta-representation $pred_a$ of anchor a instead of $factor_a$ in 16, eventually Eq. (17) (Asymmetric Triplet Loss) derived. We deem each triplet (x_a, x_p, x_n) ((x_{c_i}, C_i) while for camera-wise assignment gate) as a metric meta, and the meta-metric establishes relative positional relationships based on the anchor as a reference. The quality of meta-metric convergence is closely related to the representation quality of the anchor.

Analysis. For different meta-representations during the same training stage, the probability of correctly predicting the true class varies, reflecting the model's differing levels of fit for different samples. A higher probability of predicting the true class indicates that the sample has high label purity, while a lower probability suggests that the sample is of lower quality or has high noise. For high-purity samples, we strengthen the learning intensity, while the influence of low-quality, high-noise samples is weakened. For the same meta-representation at different training stages, the probability of correctly predicting the true class also varies. Moreover, as $pred_a$ increases, it effectively reduces the triplet loss margin α. This shift means that the model transitions from pursuing coarse-grained inter-class separability to focusing on fine-grained intra-class compactness and semantic correlations between embeddings.

$$L_{AsyT} = max(0, pred_a \cdot d_{ap} - pred_a \cdot d_{an} + \alpha) \quad (17)$$

To facilitate the model's learning and control of the scale of d, we use the exponential function and introduce a temperature coefficient constant into the exponent. Thus, $pred_a$ is rewritten as $Pred_a$:

$$Pred_a = e^{\tau \cdot pred_a} : pred_a = \lambda_1 \cdot P_i^i + \lambda_2 \quad (18)$$

Finally, Our model's loss is composed of:

$$L = L_{Cls} + L_{AysM} + \gamma L_{AysT} \quad (19)$$

L_{Cls}, L_{AysM} and L_{AysT}. The more general formula for MetaRep is:

$$L_{MR}(x_i, x_j) = Pred(x_i) \cdot D(f(x_i), f(x_j)) \quad (20)$$

3.3. The Concept of MetaRep (main) and MBN

MetaRep and Mutual Information: The meta-representation reflects the classification probability $P(Y = \hat{Y})$, representing the likelihood of the model's prediction $\hat{y}$ matching the true class y . Mutual Information $I(Y; \hat{Y})$, which quantifies the shared information between the true label Y and the prediction $\hat{Y}$, is expressed as:

$$I(Y; \hat{Y}) = H(Y) - H(Y|\hat{Y}) \quad (21)$$

Here, $H(Y)$ is the entropy of the true labels, representing their inherent uncertainty, whereas $H(Y|\hat{Y})$ quantifies the residual uncertainty conditioned on the model predictions. Treating $H(Y)$ as a constant Δ and ignoring it, the expression is simplified to:

$$I\left(Y; \hat{Y}\right) = -H\left(Y|\hat{Y}\right) = \mathbb{E}_{p(Y,\hat{Y})}\left[\log p(Y|\hat{Y})\right] \quad (22)$$

If the prediction is almost correct (i.e., $P(Y = \Omega, \hat{Y} = \Omega) \approx 1$, then $P(Y, \hat{Y}) \approx P(\hat{Y} = Y) \approx 1$. Also, $P(Y|\hat{Y}) = \frac{P(Y,\hat{Y})}{P(\hat{Y}=\Omega)} \approx \frac{1}{P(\hat{Y}=\Omega)}$. Substituting this into the mutual information expression, we obtain:

$$\begin{aligned} I\left(Y; \hat{Y}\right) \approx & P\left(\hat{Y} = Y\right) logP\left(\hat{Y}\right) \\ \approx & - L_{CE} \propto P\left(\hat{Y}\right) \end{aligned} \quad (23)$$

L_{CE} represents the cross-entropy loss. As $P(\hat{Y} = Y)$ approaches 1, the mutual information $I(Y; \hat{Y})$ becomes higher, indicating that the model has captured more information consistent with the true label. In fact, the variant of cross-entropy commonly used in contrastive learning is the InfoNCE (Information Noise-Contrastive Estimation) loss:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\sin\left(x_i, y_i\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\sin\left(x_i, y_j\right)/\tau\right)}\right] \approx L_{CE} \quad (24)$$

In 24, the numerator represents the correlation measure between the positive sample pair (x, y), while the denominator sums the correlation measures between the current sample and all other samples (for normalization). Through transformation, this ratio corresponds to the probability

of correctly clustering positive samples. Minimizing this loss function is fundamentally equivalent to maximizing the correct clustering probability of positive samples (see Eq. (25)), a process that information-theoretically equates to maximizing mutual information.

$$-\mathcal{L}_{\text{InfoNCE}} \propto P(\hat{Y} = Y) \propto I(Y; \hat{Y}) \tag{25}$$

Curriculum Learning: Through the training process, MetaRep's representations progressively improve in accuracy, as evidenced by the increase in correct class probability and mutual information, ultimately reaching a stable state. Concurrently, the step sizes of the meta-metric learning initially increase and then gradually stabilize, maintaining a consistent learning intensity. This behavior aligns with the principles of curriculum learning, which adopts a progressive difficulty scheduling approach to enhance training efficiency and effectiveness.

Incremental Learning: MBN preserves pedestrian features while removing camera-induced background biases. Its asymmetric alignment gate leverages converged representations (historical knowledge) to dynamically adjust metric alignment when adapting to new data, ensuring seamless integration of old/new domains.

Asymmetric Metric: Representation-based indicators (weighted by mutual information) enable adaptive learning paces, creating non-uniform metrics. This contrasts with symmetric metric distances:

$$d(x_i, x_j) = d(x_j, x_i) \tag{26}$$

changes to the asymmetric:

$$d^{'}(x_i, x_j) = pred_i \cdot d(x_i, x_j) \neq d^{'}(x_j, x_i) \tag{27}$$

Multi-task Convergence: MetaRep establish direct communication between metric and representation learning. The meta-representation represents the probability of correct classification in a binary classification (correct or incorrect), which reflects the mutual information between the model parameters and the sample input. However, it is inversely proportional to the relative entropy of representation learning, while the distance of the meta-metric is directly proportional to the metric learning loss. Thus, when the meta-representation and meta-metric are multiplied together as a factor, they form a negative feedback mechanism between the representation loss and the metric loss. When the model optimizes the meta-representation factor and the meta-metric factor, and backpropagates for gradient computation:

$$\frac{\partial L_{MR}}{\partial Pred} = D(f(x_i), f(x_j)) \tag{28}$$

$$\frac{\partial L_{MR}}{\partial D} = Pred(x_i) \tag{29}$$

Specifically, for MetaRep, the meta-representation factor Pred in this paper is Eq. (18). As for the meta-metric factor, in MetaRep and the camera-wise module, they are respectively distance sets in Eqs. 10 and 1: In MetaRep's framework, the co-optimization of the meta-representation and meta-metric factors establishes a dynamic feedback mechanism. When optimizing the meta-metric factor, a higher-confidence meta-representation amplifies the step size, prioritizing metric learning for high mutual information (high-quality) samples. Conversely, when optimizing the meta-representation factor, a larger meta-metric (indicating greater deviation from the positive cluster center) increases the step size, but the inverse relationship between meta-representation and classification loss counteracts this growth, reducing the classification loss step size. This creates an adaptive mechanism where samples farther from the positive cluster center incur larger classification losses, while closer samples have smaller losses. The resulting negative feedback between representation and metric losses enables consistent multi-objective convergence.

4. Experiments

4.1. Implementation

The datasets cover a diverse range of scenarios (details in Appendix), including Market1501, MSMT17, DukeMTMC, and CUHK03. We use ResNet50 pre-trained on ImageNet as the backbone. To demonstrate our method's effectiveness, we tested it on CBN and LightMBN baselines, focusing on fair ablation studies rather than engineering tricks. Evaluation metrics include mAP and CMC curves, with optimal results highlighted in bold.

CBN baseline maintains the default configuration [27]. For LightMBN, input images are normalized and resized to 384×128. Data augmentation includes resizing, random cropping, erasing, and horizontal flipping ($p = 0.5$). Training runs for 120 epochs, with a batch size of 64 (4 samples from 16 identities per batch). Adam optimizer is used with $\alpha = 1e^{-8}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Baseline losses include softmax and TriHard Loss, with MBN and MetaRep added for further validation.

4.2. Ablation of the M-MBNNet

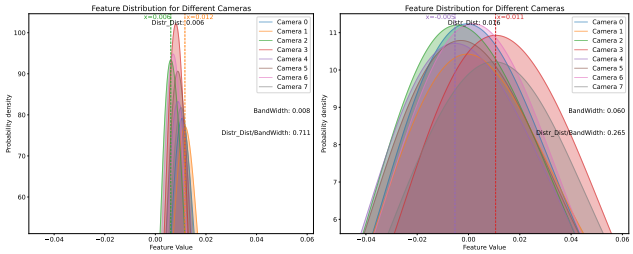

Figure 5. Visualization of camera-domain variance with /without MBN. To maintain a consistent scale, the x-axis was fixed.

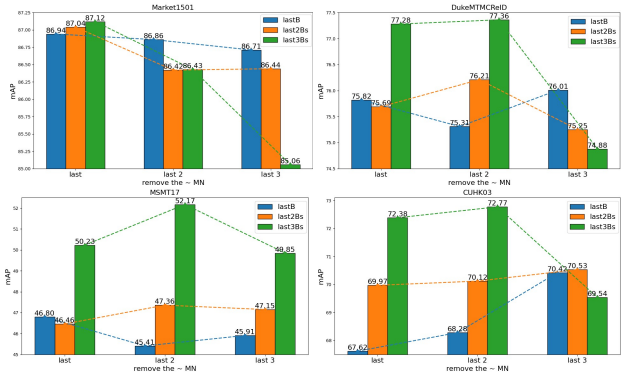

Figure 6. Ablation study on MN (Multi-Aggregation Norm). we replace the BatchNorm (BN) in ResNet50 with MN. However, since the normalization layer is influenced by the scale of the dataset, it is necessary to discard MN in certain layers with smaller statistical scales. Therefore, we conduct ablation experiments by removing the last one, last two, and last three MN blocks in layer4, as well as the last one, two, and three MN blocks in the final layers.

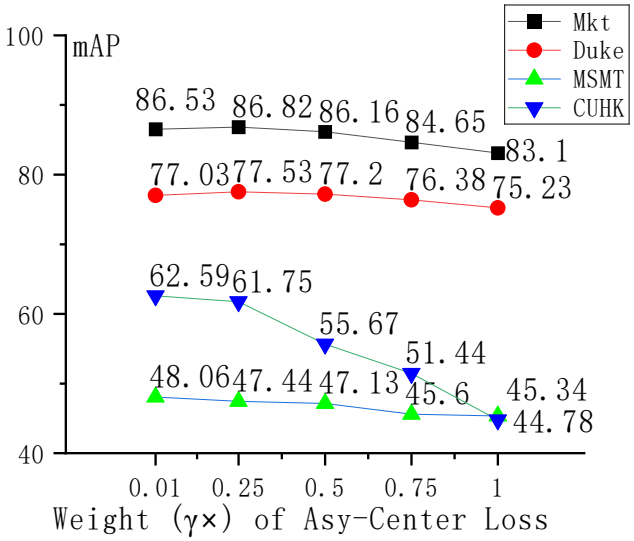

Figure 7. Weight ablation of Asy-Center Loss (of CAG).

First, we integrate MBN into the CBN (in Appendix Tab.A.5) and LightMBN baselines (Alation of CAG in Fig. 7, MN in Fig. 6) and compare the performance before and after the integration in fully supervised learning. Additionally, We further visualize the camera-domain gap before and after the integration, as shown in Fig. 5. The y-axis origin was set at the bandwidth point corresponding to the highest peak (y value: peak × 0.707). The results show that MBN reduces the domain gap, aligning all domains into a unified common space. Compared to the baseline, the maximum domain gap is reduced to 1/4, and the bandwidth is reduced to 1/6, indicating a more compact distribution. The increase in y-values (equivalent to probability density) further confirms a more concentrated distribution.

Secondly, MetaRep serves as a direct bridge connecting representation learning and metric learning, and is evaluated. It is worth mentioning that this method is also highly effective for direct transfer and semi-supervised tasks, as meta-representation can effectively leverage the knowledge from the previous training phase.

Thirdly, We gradually integrate MBN and MetaRep into the LightMBN baseline to validate the effectiveness of our approach (detail in Tab. 1). The experimental results show an average improvement of 5.5

Table 1. Ablation of M-MBNNet. Mkt represents market1501.

MBN		MR	Datasets			
CAG	MN		Mkt	Duke	MSMT17	CUHK03
			86.2|94.3	76.8|87.1	47.1|65.8	61.2|63.3
√			86.8|95.0	77.5|88.2	48.1|67.3	62.6|65.1
	√		87.1|94.5	77.4|87.7	52.2|69.9	72.8|75.4
		√	87.4|94.6	77.6|87.8	47.9|66.5	62.7|64.6
	√		87.6|95.1	78.0|88.7	52.5|70.0	73.2|75.3
	√	√	**88.1**|95.4	**78.9**|88.7	**52.9**|70.3	**73.5**|75.5

Last, to further demonstrate the potential of our method in weakly supervised and transfer learning tasks, we conducted direct transfer experiments and compared the results with the baseline. The experiments show the effectiveness of our method, as weak supervision closely approximates direct transfer followed by fine-tuning. Since this is highly consistent with the direct transfer task, we will not elaborate further. Additional ablations are provided in the appendix.

Table 2. Alation of M-MBNNet for direct transferring. Training and Testing use different data.

Trainset	TestSet	Mkt	Duke	MSMT17	CUHK03
	Method	mAP|Rank1			
market	base	86.2|94.3	14.9|27.7	1.6|4.9	2.6|2.8
	ours	88.1|95.4	34.1|53.1	3.2|9.4	7.9|7.5
Duke	base	21.9|47.8	76.8|87.1	3.1|9.2	4.6|4.3
	ours	32.5|60.1	79.0|88.7	4.8|13.8	11.5|11.7
MSMT	base	28.5|55.3	34.2|50.7	47.1|65.8	10.7|10.5
	ours	41.9|68.4	48.0|68.4	52.9|70.3	23.2|23.9
cuhk03	base	18.8|40.8	9.8|20.3	1.1|3.7	61.2|63.3
	ours	30.5|55.1	19.7|37.3	2.0|6.7	73.5|75.5

5. Conclusion

We propose M-MBNNet, which integrates MBN and MetaRep. MBN addresses domain shifts caused by varying camera backgrounds, normalizing data as if from a single background. MetaRep leverages meta-representation factors to guide metric learning, utilizing prior knowledge (e.g., mutual information or quality measures) to dynamically adjust category margins and ensure consistent convergence. This approach significantly enhances supervised learning for ReID tasks, achieving an average improvement of 5.53, and direct transfer learning, with an average gain of 9.09 for mAP.

References

[1] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian conference on computer vision*, pages 1142–1160. 3

[2] Jianyang Gu, Kai Wang, Hao Luo, Chen Chen, Wei Jiang, Yuqiang Fang, Shanghang Zhang, Yang You, and Jian Zhao. Msinet: Twins contrastive search of multi-scale interaction for object reid. 2023. 2, 3

[3] Nianchang Huang, Kunlong Liu, Yang Liu, Qiang Zhang, and Jungong Han. Cross-modality person re-identification via multi-task learning. *Pattern Recognition*, 128:108653, 2022. 1, 3

[4] Jianjun Lei, Tianyi Qin, Bo Peng, Wanqing Li, Zhaoqing Pan, Haifeng Shen, and Sam Kwong. Reducing background induced domain shift for adaptive person re-identification. *IEEE Transactions on Industrial Informatics*, 19(6):7377–7388, 2022. 1

[5] 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. 3

[6] Zongyi Li, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Boyuan Liu, Runsheng Wang, and Chengxin Zhao. Viewpoint disentangling and generation for unsupervised object re-id. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5):1–23, 2024. 3

[7] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 80–88. 3

[8] Min Liu, Fei Wang, Xueping Wang, Yaonan Wang, and Amit K Roy-Chowdhury. A two-stage noise-tolerant paradigm for label corrupted person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4944–4956, 2024. 2, 3

[9] Yang Liu, Hao Sheng, Shuai Wang, Yubin Wu, and Zhang Xiong. Feature-level camera style transfer for person re-identification. *Applied Sciences*, 12(14):7286, 2022. 1

[10] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019. 4

[11] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Ag-reid. v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2896–2908, 2024. 1

[12] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. A simple and light-weight attention module for convolutional neural networks. *International journal of computer vision*, 128(4):783–798, 2020. 3

[13] Yunjie Peng, Jinlin Wu, Boqiang Xu, Chunshui Cao, Xu Liu, Zhenan Sun, and Zhiqiang He. Deep learning based occluded person re-identification: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–27, 2023. 1

[14] Dang H Pham, Anh D Nguyen, and Hoa N Nguyen. Gan-based data augmentation and pseudo-label refinement with holistic features for unsupervised domain adaptation person re-identification. *Knowledge-Based Systems*, 288:111471, 2024. 2

[15] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206. 1

[16] Kaijie Ren and Lei Zhang. Implicit discriminative knowledge learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 393–402. 1

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823. 3

[18] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification, 2018. 2

[19] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022. 3

[20] 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 (ECCV)*, pages 480–496. 3

[21] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*, pages 499–515. Springer. 2

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19. 3

[23] Xin Xu, Xin Yuan, Zheng Wang, Kai Zhang, and Ruimin Hu. Rank-in-rank loss for person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2s):1–21, 2022. 3

[24] Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17117–17126. 3

[25] Yulin Zhang, Bo Ma, Yuqing Feng, and Meng Li. Pmtnet: Progressive multi-task network for one-shot person re-identification. *Information Sciences*, 568:133–146, 2021. 3

[26] Xiangping Zhu, Xiatian Zhu, Minxian Li, Pietro Morerio, Vittorio Murino, and Shaogang Gong. Intra-camera supervised person re-identification. *International journal of computer vision*, 129:1580–1595, 2021. 2

[27] Zijie Zhuang, Longhui Wei, Lingxi Xie, Haizhou Ai, and Qi Tian. Camera-based batch normalization: An effective distribution alignment method for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):374–387, 2021. 2, 7