Noname manuscript No.
(will be inserted by the editor)

Robust Object Re-identification with Coupled Noisy Labels

Mouxing Yang · Zhenyu Huang · Xi Peng

Received: date / Accepted: date

Abstract In this paper, we reveal and study a new challenging problem faced by object Re-IDentification (ReID), *i.e.*, Coupled Noisy Labels (CNL) which refers to the Noisy Annotation (NA) and the accompanied Noisy Correspondence (NC). Specifically, NA refers to the wrongly-annotated identity of samples during manual labeling, and NC refers to the mismatched training pairs including false positives and false negatives whose correspondences are established based on the NA. Clearly, CNL will limit the success of the object ReID paradigm that simultaneously performs identity-aware discrimination learning on the data samples and pairwise similarity learning on the training pairs. To overcome this practical but ignored problem, we propose a robust object ReID method dubbed Learning with Coupled Noisy Labels (LCNL). In brief, LCNL first estimates the annotation confidences of samples and then adaptively divides the training pairs into four groups with the confidences to rectify the correspondences. After that, LCNL employs a novel objective function to achieve robust object ReID with theoretical guarantees. To verify the effectiveness of LCNL, we conduct extensive experiments on five benchmark datasets in single- and cross-modality object ReID tasks compared with 14 algorithms. The code could be accessed from https://github.com/XLearning-SCU/2024-IJCV-LCNL.

Keywords Person Re-Identification · Cross-modality Person Re-Identification · Vehicle Re-Identification · Noisy Labels · Noisy Correspondence

Corresponding author: Xi Peng.

M. Yang, Z. Huang, and X. Peng
College of Computer Science, Sichuan University. Chengdu, China.
E-mail: {yangmouxing, zyhuang.gm, pengx.gm}@gmail.com

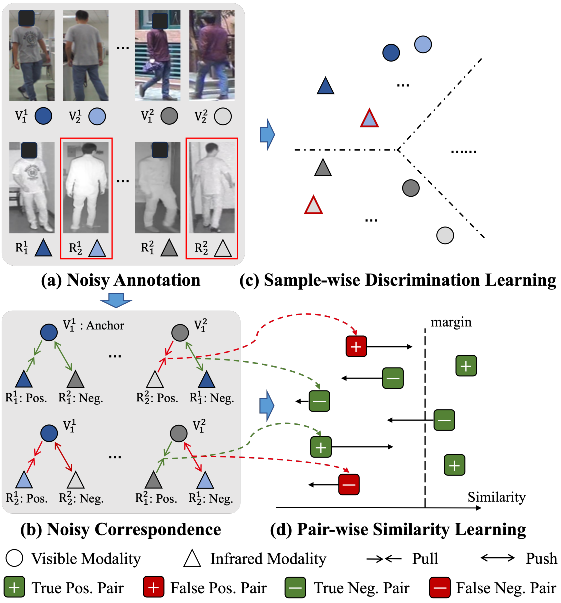


Fig. 1 Our observation on Coupled Noisy LabeLs (CNL) problem. CNL refers to the noisy annotation and the accompanied noisy correspondence. (**a**) Noisy Annotation (NA): it refers to the wrong annotations of samples. (**b**) Noisy Correspondence (NC): it refers to the mismatched pairs including false positive and false negative pairs. Without loss of generality, taking the cross-modality ReID as an example, two samples R_2^1 and R_2^2 with similar poses, which should be of identity 2 and 1 respectively. Due to the over-high similarity, however, they are probably been wrongly annotated and such a wrong annotation will further result in the false correspondence because the object pairs are usually established based on annotations. In the figure, V/R denotes the visible/infrared modality, and V_j^i denotes the j-th samples of the i-th person.

1 Introduction

For a given query, object Re-IDentification (ReID) (Bai et al., 2017; Ge et al., 2020; He et al., 2021; Luo et al., 2022; Rao et al., 2021; Ye et al., 2021b; Zheng et al., 2012, 2015b) aims at searching different images of the same identity from the gallery set, which plays an important role in the intelligent surveillance system. In the heart of ReID, the key is matching a specified object across non-overlapping visible cameras, which is generally formulated as a single-modality matching problem. Although single-modality ReID has achieved promising performance in a number of scenarios, it cannot achieve encouraging results at night due to the degraded performance of the visible camera under low illumination conditions. As a remedy, cross-modality ReID (Choi et al., 2020; Lu et al., 2020; Shi et al., 2023; Tian et al., 2021; Wu et al., 2017, 2021; Ye et al., 2021a) associates the identities across visible and infrared modalities so that the powerful capacity of infrared cameras under low-lighting conditions is exploited. Regardless of the difference in data resources, most single- and cross-modality ReID methods (Choi et al., 2020; Ge et al., 2020; He et al., 2021; Lu et al., 2020; Rao et al., 2021; Ye et al., 2021a,b; Zheng et al., 2022) share the same technical characteristics. Namely, both of them will learn the identity-aware discrimination from the annotated samples, while learning the pairwise similarity from the training pairs whose correspondences are established based on the annotations. As a result, the success of both single- and cross-modality ReID will heavily rely on the quality of data annotations.

Unfortunately, in practice, it is expensive and even impossible to precisely annotate all the samples due to the viewpoint differences across cameras, poor recognizability in the colorless infrared modality, and so on. As shown in Fig. 1(a), the analogous human poses and low image resolution probably result in Noisy Annotation (NA) which will degrade the performance of object ReID in two aspects. On the one hand, the sample-wise discrimination learning (Fig. 1(c)) will fit NA, and thus optimizing ReID models in a wrong direction. On the other hand, as almost all existing object ReID methods construct the training pairs using data annotations, NA will result in another kind of label noise, *i.e.*, Noisy Correspondence (NC, Fig. 1(b)). As shown in Fig. 1(d), the pairwise similarity learning with NC would wrongly increase the similarities of false positive pairs (FP) while decreasing the ones of false negative pairs (FN), thus degrading the performance of ReID models.

Based on the above observations, we reveal and study the Coupled Noisy Labels (CNL) problem for object ReID tasks in this paper. Note that, some recent efforts (Ge et al., 2020; Ye and Yuen, 2020; Ye et al., 2022; Yu et al., 2019) have been devoted to achieving robust ReID by generating pseudo annotations or revising noisy annotations. However, almost all of them only focus on achieving robustness on NA while ignoring the influence of NC. In fact, it is impossible to eliminate the influence of CNL by only achieving robustness against NA. To be specific, the ReID dataset usually consists of thousands of identities (categories), thus hindering the accurate revisions on NA. The inaccurate revisions on NA would still introduce the NC, which finally degrades the performance. To verify the above claims, some empirical studies will be carried out in our experiments.

To conquer the above CNL problem in ReID, we propose a robust object ReID framework, named Learning with Coupled Noisy Labels (LCNL), which could be generalized to single and cross-modality scenarios. Specifically, LCNL first models the annotation confidences by resorting to the memorization effect of Deep Neural Networks (DNNs) (Arpit et al., 2017), *i.e.*, DNNs will first fit the clean data and then noisy ones. Based on the estimated confidences, LCNL takes an adaptive way to divide training pairs into different triplet combinations with rectified correspondences, *i.e.*, True Positive pairs (TP) & True Negative pairs (TN), TP&FN, FP&TN, and FP&FN. Finally, to achieve robust ReID, LCNL adopts a novel CNL-robust objective function which consists of soft identification loss and adaptive quadruplet loss. In detail, the soft identification loss has an incentive to penalize NA by utilizing the estimated confidences. Besides, we propose an adaptive quadruplet loss which adaptively changes the optimization directions when encountering different triplet combinations, thus enjoying robustness against NC. Thanks to our loss, LCNL takes different optimization properties *w.r.t.* different homogeneous combinations (*i.e.*, TP&FN or FP&TN), which is theoretical provable. In summary, the contributions and novelties of this work are given as follows:

- We reveal a new problem faced by both single- and cross-modality object re-identification, termed coupled noisy labels. Different from existing studies on noisy annotation, CNL refers to the noise in the identities (categories) of samples and the accompanied noise in the correspondence of training pairs. To the best of our knowledge, the existing robust ReID methods only take the NA problem in single-modality person ReID into consideration. There are few studies on NA for cross-modality ReID so far, not to mention the more challenging and practical CNL problem.
- To solve the CNL problem, we propose a robust object ReID method (*i.e.*, LCNL) which enjoys the robustness against CNL for both single- and cross-modality object ReID tasks. The major novelty of LCNL is the CNL-robust object function which prevents models from CNL-dominated optimization in

two aspects. On the one hand, it achieves robustness against NA by penalizing samples of NA based on the estimated confidences. On the other hand, it achieves robustness against NC by adaptively changing optimization directions and handling homogeneous combinations with theoretical guarantees.

- Extensive experiments have been conducted on three different ReID tasks including single-modality person/vehicle ReID and cross-modality person ReID, which show the importance of the CNL problem and the effectiveness of the proposed LCNL method.

2 Related Works

In this section, we briefly review three topics related to this work, *i.e.*, deep object ReID, ReID with noisy annotations, and learning with noisy labels.

2.1 Deep Object ReID

As the two most popular tasks of object ReID, person ReID and vehicle ReID aim to match person and vehicle across cameras, respectively. In general, person ReID (He et al., 2021; Li et al., 2021; Shen et al., 2018; Suh et al., 2018; Wu et al., 2017, 2020; Ye et al., 2021a; Zheng et al., 2017b) could be roughly grouped into single- and cross-modality retrieve tasks. In brief, the single-modality person ReID aims at learning identity-aware discrimination by enlarging the inter-identity differences and alleviating the intra-identity variations caused by viewpoint differences or pose changes. According to the differences in feature learning, most of the single-modality person ReID works could be roughly grouped into the following two categories: i) the global feature learning based methods (Li et al., 2021; Wang et al., 2016; Ye et al., 2021b; Zheng et al., 2017a) which extract the global embedding for each person image by designing effective backbones or devising enhanced attention schemes; ii) the local feature learning methods (He et al., 2021; Hou et al., 2019; Sun et al., 2018) which learn part or region aggregated features to discover the nuances between different identities through image division or human parsing techniques.

Thanks to the complementarity between visible and infrared modalities, cross-modality person ReID has attracted increasing attention from the community. The greatest challenge of this task lies in how to alleviate the modality discrepancy caused by heterogeneous visible and infrared cameras. To address the challenge, a number of visible-infrared person ReID methods have been proposed, which could be classified into the following three categories, *i.e.*, i) the architecture design based methods (Choi et al., 2020; Lu et al., 2020; Wu et al., 2017, 2021; Ye et al., 2020) which strive to learn the discriminative representation shared across modalities; ii) the metric design based methods (Ye et al., 2018, 2021a,b) which aim to devise different metrics or loss functions for learning cross-modality similarity; iii) the modality transform based methods (Hao et al., 2021; Wang et al., 2019a,b; Wei et al., 2021) which aim at designing transformation or augmentation strategies to narrow the gap between modalities.

Similar to person ReID, vehicle ReID owns a broad range of demands in intelligent transportation surveillance systems. Thanks to the development of different vehicle benchmarks (Liu et al., 2016a,b, 2017), vehicle ReID has achieved promising progress during past years, which could be partitioned into two groups according to the usage of extra viewpoint information. In brief, the viewpoint-aware based methods (Chen et al., 2020; Chu et al., 2019; He et al., 2021; Meng et al., 2020) usually utilize the available orientation information to eliminate scene bias and learn invariant features. Besides, the other group of methods (Rao et al., 2021; Zhang et al., 2020) tries to distinguish the fine-grained visual differences between vehicles for enlarging the intra-identity similarity while shortening the inter-identity one.

Although huge success has been achieved during past years in the ReID community, most of the existing methods might suffer from performance degradation in some scenarios. To be specific, almost all existing ReID methods assume that the identity annotations are faultless and the training pairs are correctly matched. However, either of the two assumptions is hard and even impossible to be satisfied in real-world applications due to the extremely-large identity number and the complex data collection environment. Therefore, the existing ReID methods probably show inferior performance when encountering the coupled noisy labels as discussed in Introduction. To achieve robustness against CNL, this study formally reveals the CNL problem and proposes a CNL-robust framework for single- and cross-modality ReID. To the best of our knowledge, this study could be one of the first works on CNL-robust ReID.

2.2 Robust Object ReID with Noisy Annotations

With the rapid development of deep ReID, some works (Ge et al., 2020; Ye and Yuen, 2020; Ye et al., 2022; Yu et al., 2019) have realized the noisy annotation challenge in single-modality person ReID and a number of methods have been proposed to achieve robustness against the NA. In brief, Yu et al. (2019) first studies the NA problem in person ReID and proposes modeling the feature uncertainty to alleviate the negative impacts of noisy samples. Ye and Yuen (2020); Ye et al. (2022) aim to achieve NA-robust person ReID by explicitly correcting the annotation with model prediction. Ge et al. (2020) dives into the study of domain adaption on person ReID and proposes handling the noise during the adaption process with a carefully-designed pseudo label generating strategy.

The major differences between existing NA-robust ReID methods and this work are given below. First, existing works only consider the sample-wise NA problem for single-modality person ReID. The achieved robustness is suboptimal for the popular ReID paradigm (Choi et al., 2020; Ge et al., 2020; He et al., 2021; Lu et al., 2020; Rao et al., 2021; Ye et al., 2021a,b) which simultaneously performs sample-wise discrimination learning and pair-wise similarity learning. In contrast, this study reveals the more pragmatic CNL challenge for ReID tasks and simultaneously achieves robustness against the NA and NC, *i.e.*, CNL. Notably, there are few studies on NA for cross-modality ReID, not to mention the CNL challenge. Second, to solve the NA problem, the existing works mainly focus on revising the annotations, which is daunting for the ReID datasets of numerous identities. In contrast, our method enjoys robustness against the NA by estimating the annotation confidences and designing robust loss, which is more accessible than the explicit annotation revision. Notably, this study is also significantly different from the preliminary conference version (DART (Yang et al., 2022a)) in the following aspects. On the one hand, DART focuses on achieving the noise-robust cross-modal person ReID, whereas this work proposes a unified CNL-robust framework which could be generalized to both single- and cross-modality ReID tasks. On the other hand, the loss functions are different and the experimental studies show the superiority of this study. More specifically, DART would achieve sub-optimal robustness against the homogeneous combinations, whereas this work theoretically improves the robustness of the loss function by designing different recast functions to transform the similarities of the homogeneous combinations into desirable ones.

2.3 Learning with Noisy Labels

During the past decade, the efforts on learning with noisy labels have concentrated on the classification task (Song et al., 2020). According to the robustness paradigm, the existing studies on label noise could be roughly divided into four groups, *i.e.*, i) robust loss based methods (Kim et al., 2021; Ma et al., 2020) which aim to design the noise-tolerant loss functions; ii) robust architecture based methods (Goldberger and Ben-Reuven, 2017; Xiao et al., 2015) which modify the network architecture to estimate the noise transition matrix; iii) sample selection based methods (Han et al., 2018) which select truly-labeled data from the noisy dataset for better optimization; iv) Semi-supervised learning based methods (Li et al., 2020; Nguyen et al., 2019) which partition the dataset into clean and noisy subsets which are fed into semi-supervised learning methods (Berthelot et al., 2019). Besides the noise-robust classification studies, some recent efforts (Hu et al., 2021; Mandal and Biswas, 2020) have been devoted to solving the label noise problem for the cross-modal retrieval task.

Among the aforementioned works, the sample selection based methods and noise-robust cross-modal retrieve studies could be most similar to this work while being with the following differences. First, traditional label noise studies mainly focus on the sample-wise annotation errors. In contrast, this work consider a new label noise paradigm, *i.e.*, CNL, which refers to both sample-wise annotation errors (*i.e.*, NA) and pair-wise correspondence mismatch (*i.e.*, NC). Besides the difference in the paradigm, the proposed method is remarkably different from the sample selection based methods. In brief, the sample selection based methods usually treat the training data with relatively greater loss value as noise and discard them, which may eliminate numerous informative samples. Some of them (Han et al., 2018; Shen and Sanghavi, 2019) even require taking the noise rate as a prior. In contrast, our method first estimates the truly-annotated confidences and utilizes them to penalize the noisy samples during optimization instead of simply discarding and requiring additional priors. Besides, based on the computed confidences, this work further achieves robustness against the NC. Second, most of the robust cross-modal retrieve methods (Hu et al., 2021; Mandal and Biswas, 2020) use the off-the-shelf data pairs and assume that the training pairs are fully aligned at the instance level. In other words, they do not need to construct training pairs and assume the cross-modal correspondences are faultless. In contrast, this work dives into the object ReID task where the training pairs are constructed according to the annotations. Once the annotation is false, the NC would be inevitably introduced and the CNL-robust methods are highly expected.

2.4 Learning with Noisy Correspondence

Learning with noisy correspondence is a recently-rising topic, which mainly focuses on combating the potential mismatched pairs in cross-modal tasks. Yang

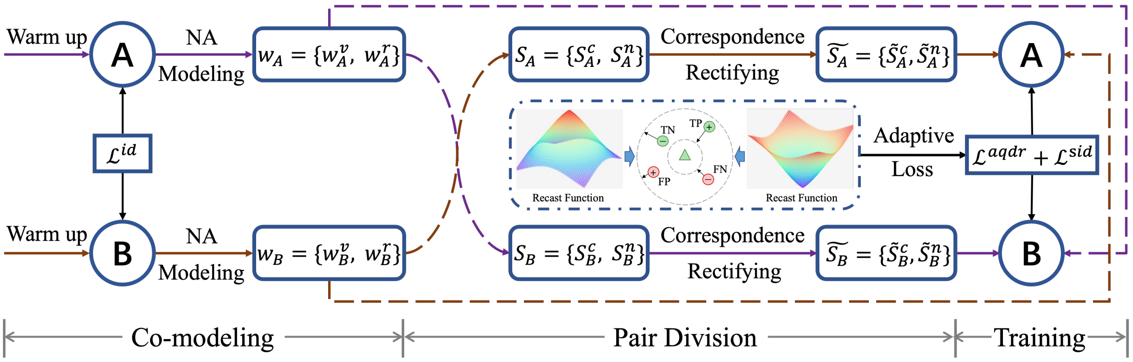


Fig. 2 The framework of LCNL. It consists of co-modeling, pair division, and dually-robust training modules. Specifically, the co-modeling module first warms up two individually-initialized DNNs with the vanilla identification loss ($\mathcal{L}^{id}$), and then models the annotation confidences by resorting to the memorization effect. After that, the estimated confidences are passed into the pair division module in a swapping manner for further usage. In the pair division module, the training pairs will be partitioned into different groups and the correspondences within each group will be rectified. As a result, four kinds of triplet combinations could be obtained, *i.e.*, TP&TN, FP&FN, TP&FN, and FP&TN. To prevent networks B/A from overfitting the NA, the estimated confidences by network A/B will be used as coefficients in the soft identification loss $\mathcal{L}^{sid}$ to penalize the noise samples. To achieve robustness against the NC, the loss $\mathcal{L}^{aqdr}$ will adaptively change the optimization directions when encountering different kinds of divided triplets. Especially, for different homogeneous triplets (*i.e.*, FP&TN or TP&FN), the loss can enjoy different optimization properties under the help of the designed recast functions, thus further improving the robustness with theoretical guarantees.

et al. (2021, 2022b) studies the false negative problem in contrastive learning and achieves robust multi-view clustering accordingly. Huang et al. (2021) first formally studies the noisy correspondence problem and achieves robust cross-modal matching against false positive pairs. Following (Huang et al., 2021), some recent works (Hu et al., 2023; Qin et al., 2022) propose solving the NC problem in more efficient and diversified ways, constantly improving the robustness and performance. Recently, some works extend the scenario of the NC problem from cross-modal matching to visible-infrared person ReID (Yang et al., 2022a) and graph matching (Lin et al., 2023). Different from the existing works, this work not only extends the definition of noisy correspondence to both false negative and false positive correspondence but also extends the setting of NC from cross-modal to both single- and cross-modal scenarios.

3 Method

In this section, we elaborate on the proposed LCNL which is a general framework for achieving robustness against the CNL encountered in both single- and cross-modality object ReID.

3.1 Problem Definition

For a given query image, most existing single- or cross-modality object ReID methods aim at finding images of the same identities within or across modalities from the gallery. For ease of representation, we take the Visible-Infrared cross-modality ReID (VI-ReID) task as a showcase without loss of generality. Formally, let $\mathcal{D}_{m_1} = \{\boldsymbol{x}_i^{m_1}, y_i^{m_1}\}_{i=1}^{N_{m_1}}$ and $\mathcal{D}_{m_2} = \{\boldsymbol{x}_i^{m_2}, y_i^{m_2}\}_{i=1}^{N_{m_2}}$ denote the observed visible and infrared modality datasets collected from K different identities respectively, where $\boldsymbol{x}_i^m$ is the image, N_m is the dataset size, $m \in \{m_1, m_2\}$ denotes the modality, and y_i^m is the identity annotation which is potentially wrong. For achieving cross-modal individual retrieve, most existing methods (Choi et al., 2020; Ge et al., 2020; He et al., 2021; Lu et al., 2020; Rao et al., 2021; Ye et al., 2021a,b) construct cross-modal set $\mathcal{S} = \left\{(\boldsymbol{x}_i^{m_1}, \boldsymbol{x}_j^{m_2}, y_{ij}^p) \mid y_{ij}^p \in \{0,1\}, i \in [1, N_{m_1}], j \in [1, N_{m_2}]\right\}$ based on the annotations, where y_{ij}^p is the pairwise correspondence indicating that the pair $(\boldsymbol{x}_i^{m_1}, \boldsymbol{x}_j^{m_2})$ is positive ($y_{ij}^p = 1$) or negative ($y_{ij}^p = 0$). In other words, $y_{ij}^p = 1$ *i.f.f.* $y_i^{m_1} = y_j^{m_2}$, and $y_{ij}^p = 0$ otherwise. With the annotated samples and constructed pairs, the methods usually adopt the sample-wise discrimination loss (*e.g.* Cross-Entropy (CE) loss) on $\mathcal{D}_m$ to learn the identity-aware discrimination, and pair-wise similarity loss (*e.g.*, triplet loss) on $\mathcal{S}$ to further enlarge the inter-identity distinguishability while alleviating the intra-identity variances.

Unfortunately, as the annotation y_i^m may be wrong due to inevitable manual labeling faults (*i.e.*, Noisy Annotation, NA), the established correspondence y_{ij}^p may also be wrong, thus leading to the so-called Noisy Correspondence (NC). Note that, the ground-truth annotations and correspondences are unknown, which are denoted as $\hat{y}_i^m$ and $\hat{y}_{ij}^p$ respectively. For simplicity, we refer to the above NA and the accompanied NC as the CNL with the following definitions.

Definition 1 Coupled Noisy Labels (CNL). For the given multi-modal dataset $\{\mathcal{D}_{m_1}, \mathcal{D}_{m_2}\}$ and the constructed cross-modal set $\mathcal{S}$, CNL means that the annotations y_i^m are of NA and the correspondences y_{ij}^p are of NC, and the ground-truth $\hat{y}_i^m$ and $\hat{y}_{ij}^p$ are agnostic.

Definition 2 Noisy Annotation (NA). For each modality $\mathcal{D}_m$, it is with NA when

$$\sum_{i=1}^{N_m} \mathbb{I}\left(y_i^m = \hat{y}_i^m\right) < N_m, \forall m \in \{m_1, m_2\}, \qquad (1)$$

where $\mathbb{I}(y_i^m = \hat{y}_i^m)$ equals to 1 *i.f.f.* $y_i^m = \hat{y}_i^m$, otherwise 0.

Definition 3 Noisy Correspondence (NC). The pairs in cross-modal set $\mathcal{S}$ consist of four types, *i.e.*, True Positive pairs (TP, $y_{ij}^p = \hat{y}_{ij}^p = 1$), True Negative pairs (TN, $y_{ij}^p = \hat{y}_{ij}^p = 0$), False Positive pairs (FP, $y_{ij}^p = 1, \hat{y}_{ij}^p = 0$) and False Negative pairs (FN, $y_{ij}^p = 0, \hat{y}_{ij}^p = 1$). NC refers to the mismatched pairs, *i.e.*, FP and FN.

Note that, the aforementioned notations and definitions are also hold for single-modality ReID cases by simply setting $m_1 = m_2$ and $i \neq j$. To achieve CNL-robust object ReID, we proposed the LCNL framework in this paper. As shown in Fig. 2, LCNL consists of co-modeling, pair division, and dually robust training modules which will be elaborated on one by one below.

3.2 Co-modeling

Some pioneer works (Arpit et al., 2017) have empirically found that DNNs are apt to fit simple patterns before fitting the complex ones, thus leading to relatively small loss values for the clean (*i.e.*, simple) samples and larger loss values for the noisy (*i.e.*, complex) samples in the initial training phase. Motivated by the so-called memorization effect of DNNs, we estimate the clean confidence of each sample by fitting the per-sample loss distribution (Huang et al., 2021; Li et al., 2020). Specifically, we first compute the per-sample identification (CE) loss of each modality by feeding $\mathcal{D}_{m_1}$ and $\mathcal{D}_{m_2}$ into the given networks respectively. Mathematically,

$$\ell_{\{F^m,C\}} = \{\ell_i\}_{i=1}^{N_m} = \left\{\mathcal{L}^{id}\left(C\left(F^m(\boldsymbol{x}_i^m)\right), y_i^m\right)\right\}_{i=1}^{N_m}, \qquad (2)$$

where $\mathcal{L}^{id}$ is the vanilla CE loss, F^m denotes the modality-specific encoder for modality m, and C denotes the shared identity classifier.

Given the above computed per-sample loss, we model the loss distribution by fitting a two-component Gaussian Mixture Model (GMM) as follows,

$$p(\ell \mid \theta) = \alpha_1 \Phi(\ell \mid \theta_1) + \alpha_2 \Phi(\ell \mid \theta_2), \qquad (3)$$

where θ denotes the parameters of GMM, $\{\theta_1, \alpha_1\}$ and $\{\theta_2, \alpha_2\}$ denote the parameter and mixture coefficient for each component, respectively. To optimize the GMM, we adopt the widely-used EM algorithm. After that, we estimate the clean confidences of annotations by computing the posterior probability of each sample belonging to the component with a small mean value based on the memorization effect of DNNs. In detail, the confidence w_i^m is computed by

$$w_i^m = p(\theta_1 \mid \ell_i) = \frac{p(\theta_1)p(\ell_i \mid \theta_1)}{p(\ell_i)}, \qquad (4)$$

where θ_1 and θ_2 denote the components with smaller and larger mean value, respectively.

The estimated annotation confidences would be utilized for NC detecting and further training. However, our empirical results show that simply training the networks in a self-modeling manner may have an incentive to accumulate errors. Hence, to circumvent the self-modeling bias, we adopt a co-modeling manner. Specifically, we individually train two sets of network $\{F_A^m, C_A\}$ and $\{F_B^m, C_B\}$ with different initializations. At each epoch, we estimate the annotation confidences for network A/B, and use them to detect NC and further train the other network B/A. Notably, as the memorization effect requires initial training to enlarge the loss value difference between clean and noisy samples, we propose warming up the two sets of networks by using the vanilla CE loss before beginning the co-modeling process.

3.3 Pair Division

Given annotation confidences estimated by the co-modeling module, the pair division module is designed to partition the cross-modal set $\mathcal{S} = \left\{(\boldsymbol{x}_i^{m_1}, \boldsymbol{x}_j^{m_2}, y_{ij}^p)\right\}$ into clean pair subset $\mathcal{S}^c$ and noisy pair subset $\mathcal{S}^n$. Formally, $\mathcal{S}^c = \left\{(\boldsymbol{x}_i^{m_1}, \boldsymbol{x}_j^{m_2}, y_{ij}^p) \mid (w_i^{m_1} \geq \gamma) \wedge (w_j^{m_2} \geq \gamma)\right\}$ and $\mathcal{S}^n = \left\{(\boldsymbol{x}_i^{m_1}, \boldsymbol{x}_j^{m_2}, y_{ij}^p) \mid ((w_i^{m_1} \geq \gamma) \wedge (w_j^{m_2} < \gamma)) \vee ((w_i^{m_1} < \gamma) \wedge (w_j^{m_2} \geq \gamma))\right\}$, where γ is the criterion threshold and is fixed to 0.5 in our experiment. Note that, the pairs with $((w_i^{m_1} < \gamma) \wedge (w_j^{m_2} < \gamma))$ will be discarded because they are both unconfident and thus cannot been correctly divided.

Given the clean subset $\mathcal{S}^c$ and noisy subset $\mathcal{S}^n$, the correspondences of pairs within the subsets would be further rectified via the following operation:

$$\widetilde{y}_{ij}^p = \mathbb{I}(y_{ij}^p \in \mathcal{S}^c) \odot y_{ij}^p, \qquad (5)$$

where $\odot$ is the xnor operator and $\widetilde{y}_{ij}^p$ is the rectified correspondence. The above operation is designed for the following goals. In brief, for the positive pairs from

$\mathcal{S}^c$, their correspondence would be rectified as $\widetilde{y}^p_{ij} = 1$, thus being regarded as true positive (TP) pairs; otherwise, they would be regarded as false positive (FP) pairs with the rectified correspondences $\widetilde{y}^p_{ij} = 0$. Likewise, for the negative pairs from $\mathcal{S}^c$, they would be treated as true negative (TN) pairs with the rectified correspondences $\widetilde{y}^p_{ij} = 0$. Notably, if negative pairs come from $\mathcal{S}^n$, they cannot be simply considered as false negative (FN) pairs since $\boldsymbol{x}_i^{m_1}$ and $\boldsymbol{x}_i^{m_2}$ may derive from different identities, *i.e.*, TN pairs. Therefore, to recall such TN pairs, we further revise the correspondences of those negative pairs from $\mathcal{S}^n$ via

$$\widetilde{y}^p_{ij} = \mathbb{I}(C(F^{m_1}(\boldsymbol{x}_i^{m_1})) = C(F^{m_2}(\boldsymbol{x}_j^{m_2}))). \tag{6}$$

With the pair division module, each training pair from $\mathcal{S}$ could be divided into one of TP, FP, TN, or FN, which is further used for training.

3.4 CNL-robust Objective Function

With the co-modeling and pair division modules, one can obtain the estimated annotation confidences of samples and the rectified correspondences of divided pairs which are combined with our CNL-robust object function to achieve robustness against the CNL. Formally, the objective function is defined as follows,

$$\mathcal{L} = \mathcal{L}^{sid} + \mathcal{L}^{aqdr}, \tag{7}$$

where $\mathcal{L}^{sid}$ and $\mathcal{L}^{aqdr}$ are designed for achieving robustness on NA and NC, respectively. In the following, we will elaborate on them one by one.

3.4.1 Robustness against NA

Given the annotation confidence w_i^m for sample $\boldsymbol{x}_i^m$ of identity k, we propose the following soft identification loss for achieving robustness against NA,

$$\mathcal{L}^{sid} = -w_i^m * \sum_{j=1}^{K} \mathbb{I}(j = k) \log p(j \mid \boldsymbol{x}_i^m). \tag{8}$$

The proposed soft identification loss owns the following merits. First, it has an incentive to penalize the noise during optimization by utilizing the confidences instead of simply discarding the noisy samples (Han et al., 2018). Besides, our loss is more feasible than the existing NA-oriented ReID methods because it need not to struggle for generating the pseudo annotations (Ge et al., 2020) or revising the wrong annotations (Ye and Yuen, 2020; Ye et al., 2022). Considering the huge identity number, it is daunting and even impossible to precisely revise the wrong annotations or predict the pseudo annotations in practice.

3.4.2 Robustness against NC

Given pairs divided by the pair division module, *i.e.*, TP, FP, TN, and FN, to alleviate the modality gap and improve the identity-level discrimination, one could construct triplet combinations and then compute the triplet loss with the hard mining strategy (Hermans et al., 2017) on them as most existing ReID methods (Ge et al., 2020; He et al., 2021; Hermans et al., 2017; Rao et al., 2021; Ye et al., 2021a,b) do. Although the vanilla triplet loss has shown its effectiveness in the object ReID community, it would suffer from the following deficiencies. First, the vanilla triplet loss can only handle the combination of TP&TN and lacks robustness against other combinations. Second, although the hard mining strategy (Hermans et al., 2017) could improve the discrimination between identities, it has an incentive to introduce more NC. In other words, in the presence of NA, the nearest negative and farthest positive samples chosen by the hard mining strategy are susceptible to be FN and FP, respectively. Therefore, it is highly expected to design a NC-robust loss which can not only handle all the possible triplet combinations *i.e.*, TP&TN, FP&FN, TP&FN, and FP&TN, but also in defense of the superiority of the hard mining strategy.

To this end, we propose the adaptive quadruplet loss. Formally, given the triplet combination $(\boldsymbol{x}_i^{m_1}, \boldsymbol{x}_j^{m_2}, \boldsymbol{x}_s^{m_2})$, the loss is in the form of

$$\mathcal{L}^{aqdr} = \mathcal{L}^{atri} + \mathcal{L}^{qdt}, \tag{9}$$

where $\boldsymbol{x}_i^{m_1}$ is the anchor sample, $\boldsymbol{x}_j^{m_2}$ and $\boldsymbol{x}_s^{m_2}$ are the corresponding hardest positive and negative samples constructed according to the annotations, *i.e.*, $y^p_{ij} = 1$ and $y^p_{is} = 0$. $\mathcal{L}^{atri}$ aims at adaptively achieving robustness on different triplet combinations and $\mathcal{L}_{qdt}$ is a quadruplet loss term. To be specific,

$$\begin{aligned}
\mathcal{L}^{atri} =& [m + (\widetilde{y}^p_{ij} \otimes \widetilde{y}^p_{is})[(-1)^{(1-\widetilde{y}^p_{ij})} d_{ij} + (-1)^{(1-\widetilde{y}^p_{is})} d_{is}] \\
&+ (\widetilde{y}^p_{ij} \odot \widetilde{y}^p_{is})(-1)^{(1-\widetilde{y}^p_{ij})} \sigma(d_{ij}, d_{is})]_+, \\
d_{ij} =& \| F^{m_1}(\boldsymbol{x}_i^{m_1}) - F^{m_2}(\boldsymbol{x}_j^{m_2}) \|_2,
\end{aligned} \tag{10}$$

where $[\cdot]_+ = \max(\cdot, 0)$, m is the margin fixed as a constant in the experiments, $\widetilde{y}^p_{ij}$ is the rectified correspondence, $\otimes$ is the xor operator, d_{ij} is the pairwise distance, and $\sigma(\cdot, \cdot)$ is the proposed recast function. Notably, the recast function is designed for keeping the merit of the hard mining strategy when encountering the homogeneous pair combinations (*i.e.*, TP&FN or FP&TN), whose principles will be elaborated on later. Clearly, learning on TP&FN or FP&TN with $\mathcal{L}^{atri}$ would only decrease or increase the pairwise

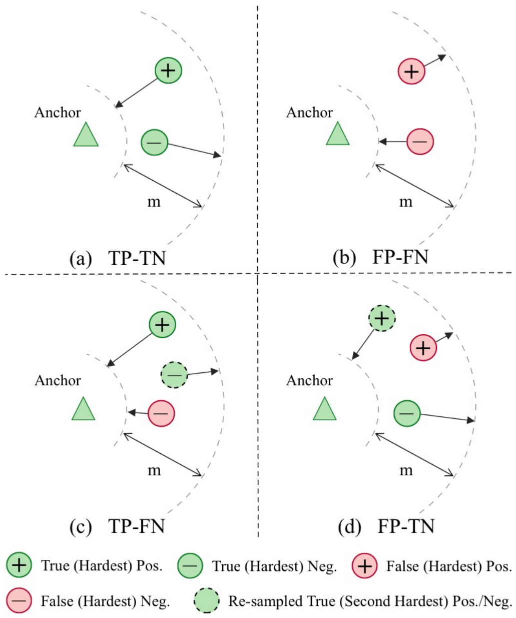


Fig. 3 The adaptive quadruplet loss could adaptively change its optimization directions for handing the four different kinds of triplet combinations, thus enjoying robustness against the NC.

distance monotonously instead of contrastively. To keep the ranking capacity of $\mathcal{L}^{atri}$, we propose the following quadruplet term,

$$\mathcal{L}^{qdt} = (-1)^{\widetilde{y}^p_{ij}\widetilde{y}^p_{is}}(\widetilde{y}^p_{ij} \odot \widetilde{y}^p_{is})d_{it}, \tag{11}$$

where d_{it} is the pairwise distance between $\boldsymbol{x}_i^{m_1}$ and $\boldsymbol{x}_t^{m_2}$, and $\boldsymbol{x}_t^{m_2}$ is the second hardest sample in the batch with confidence $w_t^{m_2} \geq \gamma$.

Besides the visual illustration in Fig. 3, we elaborate on how the proposed $\mathcal{L}^{aqdr}$ enjoys the robustness against the NC in different situations below:

- TP&TN: For the combination of TP ($y^p_{ij} = 1, \widetilde{y}^p_{is} = 1$) and TN ($y^p_{ij} = 0, \widetilde{y}^p_{is} = 0$), $\mathcal{L}^{aqdr}$ would degrade into the vanilla triplet loss which encourages to decrease and increase pairwise distances of TP and TN, respectively. Formally,

$$\mathcal{L}^{aqdr} = [m + d_{ij} - d_{is}]_+. \tag{12}$$

- FP&FN: For the combination of FP ($y^p_{ij} = 1, \widetilde{y}^p_{is} = 0$) and FN ($y^p_{ij} = 0, \widetilde{y}^p_{is} = 1$), $\mathcal{L}^{aqdr}$ would adjust its optimization tendency, *i.e.*, increasing and decreasing the distances of FP and FN, respectively.

$$\mathcal{L}^{aqdr} = [m - d_{ij} + d_{is}]_+. \tag{13}$$

- TP&FN: In this case, both TP ($y^p_{ij} = 1, \widetilde{y}^p_{is} = 1$) and FN ($y^p_{ij} = 0, \widetilde{y}^p_{is} = 1$) are positives, and thus the distance of such homogeneous pairs would be recast by $\sigma(\cdot,\cdot)$. Besides, LCNL would sample the second hardest negative sample $\boldsymbol{x}_t^{m_2}(\widetilde{y}^p_{it} = 0)$ for computing. Formally,

$$\mathcal{L}^{aqdr} = [m + \sigma(d_{ij}, d_{is}) - d_{it}]_+. \tag{14}$$

- FP&TN: Similar to the aforementioned homogeneous case, for the combination of FP ($y^p_{ij} = 1, \widetilde{y}^p_{is} = 0$) and TN ($y^p_{ij} = 0, \widetilde{y}^p_{is} = 0$), their distance would be recast by $\sigma(\cdot,\cdot)$. Then, LCNL would sample the second hardest positive sample $\boldsymbol{x}_t^{m_2}(\widetilde{y}^p_{it} = 1)$. Formally,

$$\mathcal{L}^{aqdr} = [m - \sigma(d_{ij}, d_{is}) + d_{it}]_+. \tag{15}$$

In the following, we elaborate on the principles of the proposed recast functions $\sigma(\cdot,\cdot)$ which endows $\mathcal{L}^{aqdr}$ with the hard mining capacity on the homogeneous pair combinations (FP&TN or TP&FN). Recalling that the hard mining strategy will choose the nearest negative and the furthermost positive samples as discussed above. As a result, TN in FP&TN would have a smaller distance than FP, while TP in TP&FN would have a greater distance than FN. It is expected to transform the distances of the homogeneous combination into a new one for usage in Eq. 10 while keeping the hard mining capacity on them. To this end, given the homogeneous pairs $(\boldsymbol{x}_i^{m_1}, \boldsymbol{x}_j^{m_2})$ and $(\boldsymbol{x}_i^{m_1}, \boldsymbol{x}_s^{m_2})$, we design the following five alternative recast functions σ, where $y^p_{ij} = 1, y^p_{is} = 0$, $\widetilde{y}^p_{ij} = \widetilde{y}^p_{is} = 0$ or 1. Mathematically,

$$\begin{aligned}
\sigma_1 &= \frac{d_{ij} + d_{is}}{2},\\
\sigma_2 &= \max(d_{ij}, d_{is}),\\
\sigma_3 &= \min(d_{ij}, d_{is}),\\
\sigma_4 &= \begin{cases} \max(d_{ij}, d_{is}), & \widetilde{y}^p_{ij} = \widetilde{y}^p_{is} = 1,\\ \min(d_{ij}, d_{is}), & \widetilde{y}^p_{ij} = \widetilde{y}^p_{is} = 0, \end{cases}\\
\sigma_5 &= \frac{\exp((-1)^{(1-\widetilde{y}^p_{ij})}d_{ij})}{\exp((-1)^{(1-\widetilde{y}^p_{ij})}d_{ij}) + \exp((-1)^{(1-\widetilde{y}^p_{is})}d_{is})} * d_{ij}\\
&\quad + \frac{\exp((-1)^{(1-\widetilde{y}^p_{is})}d_{is})}{\exp((-1)^{(1-\widetilde{y}^p_{ij})}d_{ij}) + \exp((-1)^{(1-\widetilde{y}^p_{is})}d_{is})} * d_{is},
\end{aligned} \tag{16}$$

Under the help of different recast functions, the NC-robust loss $\mathcal{L}^{aqdr}$ would enjoy different optimization properties when encountering the homogeneous combination TP&FN or FP&TN. To be specific, we derive the optimization properties by analyzing the gradient of $\mathcal{L}^{aqdr}$ *w.r.t.* the pairwise distance as follows.

- σ_1: $\partial\mathcal{L}^{aqdr}/\partial d_{ij} = \partial\mathcal{L}^{aqdr}/\partial d_{is} = 1/2$. The gradient value of $\mathcal{L}^{aqdr}$ *w.r.t.* d_{ij} equals to the one of $\mathcal{L}^{aqdr}$ *w.r.t.* d_{is}. As a result, the networks would learn from the homogeneous pairs equally.

- σ_2: $\partial\mathcal{L}^{aqdr}/\partial d_{ij} = 1$ and $\partial\mathcal{L}^{aqdr}/\partial d_{is} = 0$ if $d_{ij} > d_{is}$; otherwise on the contrary. The gradient is only produced with the pair of greater distance, and thus the networks will only learn from the corresponding pair.
- σ_3: Similarly, $\partial\mathcal{L}^{aqdr}/\partial d_{ij} = 0$ and $\partial\mathcal{L}^{aqdr}/\partial d_{is} = 1$ if $d_{ij} > d_{is}$; otherwise on the contrary. The gradient is only produced with the pair of smaller distance, and the networks would only learn from that pair similarly.
- σ_4: It could be regarded as the synthesis of σ_2 and σ_3. When the combination is TP&FN, σ_4 would degrade into σ_2, otherwise σ_3.
- σ_5: For the TP&FN combination, the gradient value produced with the pair of greater distance is larger; For the FP&TN combination, the gradient value produced with the pair of smaller distance is larger.

One could easily prove the properties of σ_1, σ_2, σ_3 and σ_4. For σ_5, the property on the TP&FN combination could be mathematically guaranteed by Theorem 1 and that on the FP&TN combination could be proved by Theorem 2.

Theorem 1 *For the TP&FN combination, the gradient value of $\mathcal{L}^{aqdr}$ with σ_5 w.r.t. d_{ij} is greater than that w.r.t. d_{is} when $d_{ij} > d_{is}$.*

Proof For the TP&FN combination, $\widetilde{y}^p_{ij} = \widetilde{y}^p_{is} = 1$, the gradient of $\mathcal{L}^{aqdr}$ with σ_5 *w.r.t.* d_{ij} is in the form of

$$\frac{\partial\mathcal{L}^{aqdr}}{\partial d_{ij}} = \frac{\exp(2d_{ij}) + (1 + d_{ij} - d_{is})\exp(d_{ij} + d_{is})}{(\exp(d_{ij}) + \exp(d_{is}))^2},$$

and the gradient of $\mathcal{L}^{aqdr}$ with σ_5 *w.r.t.* d_{is} is in the form of

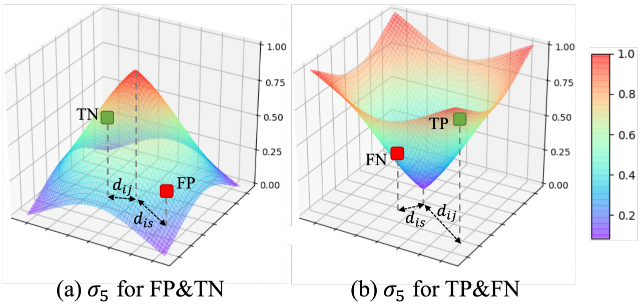
$$\frac{\partial\mathcal{L}^{aqdr}}{\partial d_{is}} = \frac{\exp(2d_{is}) + (1 + d_{is} - d_{ij})\exp(d_{ij} + d_{is})}{(\exp(d_{ij}) + \exp(d_{is}))^2}.$$

Let G be the square difference between the values of $\partial\mathcal{L}^{aqdr}/\partial d_{ij}$ and $\partial\mathcal{L}^{aqdr}/\partial d_{is}$, it could be proved that $G > 0, \forall d_{ij} > d_{is}$ by

$$\begin{aligned} G &= \left|\frac{\partial\mathcal{L}^{aqdr}}{\partial d_{ij}}\right|^2 - \left|\frac{\partial\mathcal{L}^{aqdr}}{\partial d_{is}}\right|^2 \\ &= \frac{\exp(2d_{ij}) - \exp(2d_{is}) + 2(d_{ij} - d_{is})\exp(d_{ij} + d_{is})}{(\exp(d_{ij}) + \exp(d_{is}))^2} \\ &> 0. \end{aligned}$$

Therefore, the gradient value of $\partial\mathcal{L}^{aqdr}/\partial d_{ij}$ is greater than $\partial\mathcal{L}^{aqdr}/\partial d_{is}$ when $d_{ij} > d_{is}$.

Theorem 2 *For the FP&TN combination, the gradient value of $\mathcal{L}^{aqdr}$ with σ_5 w.r.t. d_{ij} is greater than that w.r.t. d_{is} when $d_{ij} < d_{is}$.*



(a) σ_5 for FP&TN (b) σ_5 for TP&FN

Fig. 4 The gradient of $\mathcal{L}^{aqdr}$ *w.r.t.* the pairwise distances. (a) $\mathcal{L}^{aqdr}$ with σ_5 for FP&TN. (b) $\mathcal{L}^{aqdr}$ with σ_5 for TP&FN. In the figure, d_{ij} and d_{is} denote the pairwise distance of the homogeneous pairs which are both negative or positive. From the figure, one could have an intuitive understanding on Theorem 1-2.

Similarly, Theorem 2 can be proved like Theorem 1 and the details are presented in Appendix. Besides the above mathematical analysis, we also visualize the performance surfaces of $\mathcal{L}^{aqdr}$ with σ_5 in Fig. 4 for an intuitive understanding. On the one hand, Theorem 1 and 2 show that $\mathcal{L}^{aqdr}$ with σ_5 would push the pair with greater distance (*i.e.*, TP) more strongly by assigning a greater gradient, because TP&FN are both positive. Similarly, $\mathcal{L}^{aqdr}$ with σ_5 would pull the pair with smaller distance (*i.e.*, TN) more strongly for the combination TP&FN. As a result, the network optimization will be benefited.

Based on the above theoretical and visual analyses, one could have the following conclusions. To be specific, if σ_1 is used, each pair in the homogeneous combination would contribute equally to the network optimization regardless of their hardness. If σ_2 or σ_3 is used, only the pair with greater or smaller distance would contribute to the optimization. Although σ_4 enjoys the advantage of σ_2 and σ_3, it thoroughly ignores the easy pairs and thus being suboptimal. As a remedy, σ_5 could adaptively adjust the gradient according to the hardness of the pairs, thus maintaining the merits of the hard mining strategy when learning with NC. Therefore, our experiments adopt σ_5 for its effectiveness.

4 Experiments

To verify the effectiveness of LCNL for achieving CNL-robust object ReID, we conduct experiments on three different object ReID tasks including single-modality person ReID (V-ReID), vehicle ReID, and cross-modality visible-infrared person ReID (VI-ReID). The organization of this section is as follows. In section 4.1, we elaborate on the experiment settings including the hyperparameter configurations and datasets. In section 4.2, 4.3

and 4.4, we carry out quantitative and ablation studies to demonstrate the effectiveness of LCNL for achieving CNL-robust VI-ReID, V-ReID, and vehicle ReID, respectively. In section 4.5, we conduct a series of experimental analyses to show the importance of the CNL-oriented paradigm.

4.1 Settings

In this section, we elaborate on the experiment settings of LCNL including the hyper-parameter configurations and the used datasets.

Parameter Configurations: In our experiments, the warm-up epochs are fixed as 1, 5, and 10 for the VI-ReID, V-ReID, and Vehicle-ReID tasks, respectively. In addition, the margin m for the adaptive quadruplet loss and the threshold γ for the pair division module are set as 0.3 and 0.5, respectively. In the inference stage, we simply average the embeddings output by models A and B for the evaluation and no complex strategy is used. All the experiments and evaluations are conducted on Ubuntu OS with GeForce RTX 3090 GPUs.

Datasets: For the VI-ReID task, we adopt two publicly available datasets, *i.e.*, SYSU-MM01 (Wu et al., 2017) and RegDB (Nguyen et al., 2017). For the V-ReID task, we use two widely-used datasets, *i.e.*, Market-1501 (Zheng et al., 2015a) and DukeMTMC (Zheng et al., 2017b). As for the Vehicle-ReID task, we adopt the widely-used VeRi-776 (Liu et al., 2016b, 2017). Table 1 summarizes the statistics of the above five datasets and the Appendix includes more details.

Table 1 Statistics of the used datasets.

Dataset	Modality	Object	Train	Gallery	Query
SYSU-MM01	Cross	Person	34,167 (395)	301 (96)	3,803 (96)
RegDB	Cross	Person	4,120 (206)	2,060 (206)	2,060 (206)
Market-1501	Single	Person	13,387 (751)	15,913 (750)	3,368 (750)
DukeMTMC	Single	Person	16,522 (702)	17,661 (702)	2,228 (702)
VeRi-776	Single	Vehicle	37,715 (576)	11,579 (200)	1,678 (200)

For quantitative evaluations, we randomly choose a specific proportion of samples in each dataset and assign them with random identities to simulate the noisy annotations. For comprehensive investigation, the noise ratio varies from 0%, 20%, to 50%. Following Hermans et al. (2017); Liu et al. (2016b); Ye et al. (2021b); Zheng et al. (2015a, 2017b), we use two metrics for performance evaluation, *i.e.*, the mean average precision score (mAP) and the cumulative matching curve (CMC). Besides, for person ReID tasks (VI-ReID and V-ReID), we additionally use the mINP metric to measure the matching efficiency by following Ye et al. (2021b).

4.2 Robust Cross-modality Person ReID

To verify the effectiveness of LCNL, we compare LCNL with recently-published VI-ReID methods on the noisy VI-ReID datasets. In addition, we conduct ablation studies to reveal the importance of each module in LCNL on achieving robustness.

4.2.1 Comparisons with State of the Arts

To investigate the effectiveness of the LCNL framework, we employ it to endow the state-of-the-art ADP (Ye et al., 2021a) with robustness against the CNL. In the investigation, we adopt ADP's backbone and train it under the proposed LCNL framework with our CNL-robust objective function.

Following the common evaluation protocol in VI-ReID (Park et al., 2021; Wu et al., 2021; Ye et al., 2021a), we report the performance on the SYSU-MM01 dataset under the modes of "All-Search" and "Indoor-Search". For the RegDB dataset, we report the mean results of the standard 10 train/test splits under the modes of "Visible to Thermal" and "Thermal to Visible".

The performance of LCNL is compared with six recently-proposed VI-ReID methods including CrossAGW (Ye et al., 2021b), DDAG (Ye et al., 2020), LbA (Park et al., 2021), MPANet (Wu et al., 2021), ADP (Ye et al., 2021a), and DART (Yang et al., 2022a). Among them, the former five baselines are the standard VI-ReID methods, and DART is the only one that could achieve noise-robust VI-ReID. For the performance of DART, we directly take the results reported in the original paper. For the other baselines, we refer to the reported results as the ones with the noise ratio of 0% and carry out the baselines with careful parameter tuning under the noise ratio of 20% and 50%. From Table 2 and 3, one could observe that LCNL achieves stable performance while the vanilla methods encounter heavy performance degradation. Meanwhile, although LCNL is designed for achieving CNL-robust object ReID including but not limited to the VI-ReID task, it still shows promising performance improvements compared to DART which is dedicatedly designed for handling noise in VI-ReID. The above observations imply the importance of developing CNL-resistant methods for object ReID. Besides, LCNL outperforms ADP on the SYSU-MM01 dataset in terms of most metrics, even under the noise-free setting. The improvement could be attributed to that the "clean" data are probably contaminated by unrevealed noises.

Table 2 Comparisons with state-of-the-art methods on the SYSU-MM01 dataset under the noise ratio of $0\%, 20\%$ and 50%, respectively. The best and second best results are highlighted in **bold** and underline.

Noise	Methods	All-Search					Indoor-Search				
		Rank-1	Rank-10	Rank-20	mAP	mINP	Rank-1	Rank-10	Rank-20	mAP	mINP
0%	DDAG (ECCV2020)	54.8	90.4	95.8	53.0	39.6	61.0	94.1	98.4	68.0	62.6
	CrossAGW (TPAMI2021)	47.5	84.4	92.1	47.7	35.3	54.2	91.1	96.0	63.0	59.2
	LbA (ICCV2021)	55.4	–	–	54.1	–	58.5	–	–	66.3	–
	MPANet (CVPR2021)	**70.6**	96.2	98.8	**68.2**	–	**76.7**	**98.2**	99.6	**81.0**	–
	ADP (ICCV2021)	69.9	95.7	98.5	66.9	53.6	76.3	97.9	99.5	80.4	76.8
	DART (CVPR2022)	68.7	**96.4**	**99.0**	66.3	53.3	72.5	97.8	99.5	78.2	74.9
	LCNL (Ours)	70.2	**96.4**	**99.0**	68.0	**55.5**	76.2	**98.2**	**99.8**	80.3	**76.9**
20%	DDAG (ECCV2020)	14.6	46.6	61.8	14.0	5.6	15.1	50.7	69.3	22.4	18.3
	CrossAGW (TPAMI2021)	17.7	56.8	72.5	18.2	8.6	20.8	65.0	82.4	29.8	25.3
	LbA (ICCV2021)	9.9	39.5	55.9	10.2	3.8	10.1	44.1	64.5	17.4	14.0
	MPANet (CVPR2021)	21.6	63.6	78.7	21.2	–	23.8	70.2	86.4	33.2	–
	ADP (ICCV2021)	25.4	67.6	80.9	23.7	11.1	26.6	70.7	85.2	35.0	29.6
	DART (CVPR2022)	66.3	**95.3**	**98.4**	64.1	50.7	70.5	97.1	99.0	75.9	72.3
	LCNL (Ours)	**67.2**	95.1	**98.4**	**64.9**	**51.7**	**73.4**	**97.6**	**99.5**	**78.2**	**74.4**
50%	DDAG (ECCV2020)	6.7	29.0	43.8	7.5	2.9	8.4	37.9	57.9	15.1	12.3
	CrossAGW (TPAMI2021)	7.9	37.6	55.8	9.8	4.4	9.6	47.9	70.5	18.1	15.2
	LbA (ICCV2021)	2.7	17.8	30.3	4.2	1.9	4.9	29.4	49.0	11.0	8.6
	MPANet (CVPR2021)	7.0	32.8	49.2	8.2	–	8.5	40.7	61.4	15.9	–
	ADP (ICCV2021)	8.0	42.6	62.1	10.8	5.2	11.5	53.0	76.8	20.8	17.5
	DART (CVPR2022)	60.3	93.4	97.5	58.7	45.3	65.7	95.0	98.2	71.8	68.1
	LCNL (Ours)	**62.4**	**93.6**	**97.5**	**59.8**	**45.9**	**67.2**	**96.4**	**99.1**	**73.1**	**69.0**

Table 3 Comparisons with state-of-the-art methods on the RegDB dataset under the noise ratio of $0\%, 20\%$ and 50%, respectively. The best and second best results are highlighted in **bold** and underline.

Noise	Methods	Visible to Thermal			Thermal to Visible		
		Rank-1	mAP	mINP	Rank-1	mAP	mINP
0%	DDAG (ECCV2020)	69.3	63.5	49.2	68.1	61.8	48.6
	CrossAGW (TPAMI2021)	70.1	66.4	50.2	70.5	66.0	51.2
	LbA (ICCV2021)	74.2	67.6	–	72.4	65.5	–
	MPANet (CVPR2021)	83.7	**80.9**	–	82.8	**80.7**	–
	ADP (ICCV2021)	85.0	79.1	**65.3**	**84.8**	77.8	**61.6**
	DART (CVPR2022)	83.6	75.7	60.6	82.0	73.8	56.7
	LCNL (Ours)	**85.6**	78.7	65.0	84.0	76.9	60.9
20%	DDAG (ECCV2020)	39.3	25.7	10.0	37.7	25.1	9.6
	CrossAGW (TPAMI2021)	47.8	31.4	12.4	47.2	30.9	11.9
	LbA (ICCV2021)	36.0	23.5	7.5	36.2	22.8	6.7
	MPANet (CVPR2021)	33.8	23.5	–	32.6	22.1	–
	ADP (ICCV2021)	50.7	35.9	14.1	50.0	34.8	12.6
	DART (CVPR2022)	82.0	74.2	57.9	79.5	71.7	54.5
	LCNL (Ours)	**84.5**	**76.7**	**61.6**	**82.5**	**74.6**	**57.3**
50%	DDAG (ECCV2020)	24.0	14.4	4.3	21.5	13.4	4.3
	CrossAGW (TPAMI2021)	21.9	13.4	3.9	21.0	13.0	3.7
	LbA (ICCV2021)	11.7	6.7	1.5	10.2	6.3	1.5
	MPANet (CVPR2021)	9.5	6.1	–	11.4	6.7	–
	ADP (ICCV2021)	17.0	11.3	3.6	20.3	12.3	3.2
	DART (CVPR2022)	**78.2**	**67.0**	**48.4**	**75.0**	**64.4**	**43.6**
	LCNL (Ours)	76.3	65.9	47.9	73.8	63.2	42.9

Table 4 Ablation studies on SYSU-MM01 with 20% noise. The default setting is marked in gray .

Method Variants				20% noise	
Co-Modeling	Pair Division	$\mathcal{L}^{sid}$	$\mathcal{L}^{aqdr}$	mAP	mINP
✓				29.8	15.2
✓		✓		62.2	48.5
✓	✓	✓		63.8	49.8
✓	✓	✓	✓	**64.9**	**51.7**

4.2.2 Ablation Studies

To investigate the importance of each module of LCNL, we conduct ablation studies on the SYSU-MM01 dataset with 20% noise. In detail, we perform LCNL by discarding or replacing the modules and evaluating the corresponding variants. More specifically, i) we train the baseline using the co-modeling module with two individually networks; ii) we add the proposed soft identification loss ($\mathcal{L}^{sid}$) for achieving robustness on the NA; iii) we add the pair division module and replace the designed adaptive quadruplet loss ($\mathcal{L}^{aqdr}$) with the vanilla triplet loss (Hermans et al., 2017). Namely, LCNL is performed only on the clean combination (TP&TN); iv) the complete pipeline of LCNL. As shown in Table 4, it is promising to simultaneously embrace the robustness on both NA and NC (Row 4) instead of only on NA (Row 2) or neither (Row 1). Furthermore, the vanilla triplet loss, which can only handle the clean combination (*i.e.*, TP&TN) is sub-optimal for achieving robustness against the CNL (Row 3).

4.3 Robust Singe-modality Person ReID

In this section, we apply LCNL on the noisy V-ReID datasets and compare the performance with several noise-robust V-ReID methods. Besides, we conduct ablation studies to investigate the effect of each component of LCNL.

4.3.1 Comparison with State of the Arts

In this section, we endow VisibleAGW (Ye et al., 2021b) with the robustness against CNL under our LCNL framework. In brief, we adopt the backbone of VisibleAGW (Ye et al., 2021b) and our learning paradigm with the CNL-robust objective function.

The performance of LCNL is compared with five V-ReID methods including VisibleAGW (Ye et al., 2021b),

Table 5 Comparisons with state-of-the-art methods on the Market1501 and DukeMTMC datasets under the noise ratio of 0%, 20% and 50%, respectively. The best and second best results are highlighted in **bold** and underline.

Noise	Methods	Market1501					Duke-MTMC				
		Rank-1	Rank-5	Rank-10	mAP	mINP	Rank-1	Rank-5	Rank-10	mAP	mINP
0%	DistributionNet (ICCV2019)	87.3	94.7	96.7	70.8	–	74.7	85.1	88.2	56.0	–
	PurifyNet (TIFS2020)	88.4	95.8	97.6	72.1	–	77.8	88.6	92.4	62.0	–
	MMT (ICLR2020)	89.2	96.2	97.8	74.1	–	78.2	88.6	92.1	64.0	–
	VisibleAGW (TPAMI2021)	**95.1**	98.2	**99.0**	**87.8**	**65.0**	**88.9**	**95.3**	**96.7**	**79.6**	**45.7**
	CORE (TIP2022)	89.6	96.4	98.1	74.6	–	78.8	89.4	92.6	64.1	–
	LCNL (Ours)	94.7	**98.4**	**99.0**	87.7	64.5	88.7	94.7	96.4	79.3	43.1
20%	DistributionNet (ICCV2019)	77.0	90.6	94.0	53.4	–	62.4	77.4	82.5	40.9	–
	PurifyNet (TIFS2020)	83.1	93.3	95.9	63.1	–	74.1	85.6	89.2	55.8	–
	MMT (ICLR2020)	79.2	91.8	95.2	57.8	–	70.5	84.9	88.9	54.7	–
	VisibleAGW (TPAMI2021)	80.8	93.5	96.3	59.3	20.3	68.4	85.2	89.6	52.2	15.0
	CORE (TIP2022)	84.1	93.1	95.5	66.2	–	74.4	85.9	89.7	55.8	–
	LCNL (Ours)	**94.4**	**97.9**	**98.9**	**86.6**	**61.7**	**87.7**	**93.9**	**96.0**	**77.6**	**40.4**
50%	DistributionNet (ICCV2019)	61.1	81.1	87.1	35.1	–	46.0	63.9	70.9	25.8	–
	PurifyNet (TIFS2020)	83.4	94.1	96.3	52.1	–	65.0	79.0	83.9	44.5	–
	MMT (ICLR2020)	55.6	76.5	83.1	31.7	–	51.0	67.6	74.4	34.9	–
	VisibleAGW (TPAMI2021)	51.2	72.4	79.7	27.1	3.3	42.0	61.7	70.1	26.2	3.4
	CORE (TIP2022)	80.1	91.5	94.4	46.2	–	56.9	72.6	77.3	37.5	–
	LCNL (Ours)	**90.9**	**97.3**	**98.3**	**79.7**	**47.6**	**83.0**	**92.1**	**94.5**	**71.5**	**30.9**

CORE (Ye et al., 2022), MMT (Ge et al., 2020), PurifyNet (Ye and Yuen, 2020), DistributionNet (Yu et al., 2019). Among them, (Ge et al., 2020; Ye and Yuen, 2020; Ye et al., 2022; Yu et al., 2019) are NA-robust V-ReID methods, and our method is the only CNL-robust approach. As illustrated in Table 5, as the noise ratio increases, the performance of all the baselines remarkably reduces. In contrast, LCNL performs stably, which verifies that the robustness against the CNL is more favorable compared to NA. Besides the superiority of LCNL in noisy cases, it performs comparably to VisbleAGW (Ye et al., 2021b) under the noise-free setting.

Table 6 Ablation studies on Market1501 with 20% noise. The default setting is marked in gray .

Method Variants				20% noise	
Co-Modeling	Pair Division	$\mathcal{L}^{sid}$	$\mathcal{L}^{aqdr}$	mAP	mINP
✓				66.2	28.3
✓		✓		85.5	58.9
✓	✓	✓		84.5	57.0
✓	✓	✓	✓	**86.6**	**61.7**

4.3.2 Ablation Studies

In this section, we carry out ablation studies on Market-1501 with 20% noise. Similar to Section 4.2.2, we investigate the performances of different variants of LCNL. From Table 6, one could observe that each module of LCNL plays an inseparable role in achieving CNL-robust V-ReID.

4.4 Robust Vehicle ReID

In this section, we perform LCNL on the noisy Vehicle-ReID datasets compared with some state-of-the-art vehicle ReID baselines. Moreover, we also carry out ablation studies accordingly.

4.4.1 Comparisons with State of the Arts

We endow TransReID (He et al., 2021), with the robustness against CNL under our LCNL framework. In our implementation, we adopt TransReID's backbone and train it under the LCNL pipeline with the CNL-robust objective function.

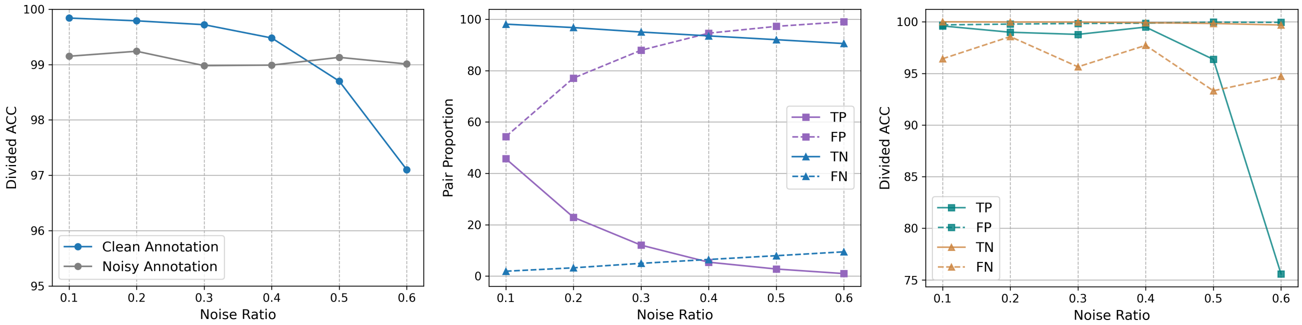
Besides the vanilla TransReID (He et al., 2021), we also compare LCNL with two recently-published Vehicle-ReID methods including PGAN (Zhang et al., 2020) and PVEN (Meng et al., 2020). As demonstrated in Table 7, LCNL substantially achieves robustness on CNL under different noise ratios while the baselines all fail.

4.4.2 Ablation Studies

In this section, we conduct ablation studies on the VeRi-776 dataset with 20% noise. Similar to Section 4.2.2, we perform LCNL with different variants and summarize the corresponding results in Table 8. From the results, one could conclude that each module of LCNL plays an important role in achieving robustness on CNL.

4.5 Analysis Experiments

To further analyze the revealed CNL problem and the proposed LCNL framework, we conduct experiments on



(a) Detection on NA (b) Statistics on NC (c) Detection on NC

Fig. 5 The ability to detect the NA and NC with different noise ratios. “Divided ACC” denotes the detection accuracy of NA / NC. “Pair Proportion” denotes the population statistics of different kinds of pairs.

Table 7 Comparisons with state-of-the-art methods on the VeRi-776 dataset under the noise ratio of 0%, 20% and 50%, respectively. The best and second best results are highlighted in **bold** and <u>underline</u>.

Noise	Methods	VeRi-776			
		Rank-1	Rank-5	Rank-10	mAP
0%	PGAN (TIFS2020)	96.5	–	–	79.3
	PVEN (CVPR2020)	95.6	**98.4**	–	79.5
	TransReID (ICCV2021)	**97.1**	–	–	**82.0**
	LCNL (Ours)	<u>96.9</u>	**98.4**	**98.9**	<u>81.7</u>
20%	PGAN (TIFS2020)	76.6	91.9	95.1	42.3
	PVEN (CVPR2020)	76.9	89.6	94.9	47.1
	TransReID (ICCV2021)	<u>83.0</u>	<u>93.9</u>	<u>96.5</u>	<u>49.7</u>
	LCNL (Ours)	**97.4**	**98.7**	**99.3**	**81.8**
50%	PGAN (TIFS2020)	34.6	59.7	70.4	10.5
	PVEN (CVPR2020)	<u>54.1</u>	<u>71.6</u>	<u>80.5</u>	<u>21.6</u>
	TransReID (ICCV2021)	49.6	69.0	77.4	13.0
	LCNL (Ours)	**96.7**	**98.3**	**99.1**	**79.4**

Table 8 Ablation studies on VeRi-776 with 20% noise. The default setting is marked in gray .

Method Variants				20% noise
Co-Modeling	Pair Division	$\mathcal{L}^{sid}$	$\mathcal{L}^{aqdr}$	mAP
✓				52.5
✓		✓		77.5
✓	✓	✓		72.6
✓	✓	✓	✓	**81.8**

the VI-ReID task with the SYSU-MM01 dataset under the all-search evaluation mode.

4.5.1 Necessity of CNL-oriented Techniques

In Introduction, we argue that it is impossible to eliminate the influence of CNL by only achieving robustness against the NA. In this section, we verify this claim by taking the cross-modality VI-ReID task as a showcase. As there is no NA-robust VI-ReID method yet, we take the generalized approaches as alternatives. Specifically, we adopt DivideMix (Li et al., 2020) for annotation rectification and then adopt the ADP (Ye et al., 2021a) on the rectified data. Furthermore, we report the performance of ADP by only using the clean data for training. As illustrated in Table 9, one could have the following observations. First, DivideMix performs imperfect data division and annotation rectification, thus leading to inferior VI-ReID performance. Second, the inferior performance of ADP-Clean compared to LCNL demonstrates that it is inadvisable to simply discard the noisy sample even using the ground truth partition as prior. The above experimental results support our claims, showing the importance of developing the CNL-oriented methods for the object ReID tasks.

Table 9 The necessity of our CNL-oriented paradigm. “DA” denotes the division accuracy on clean and noise samples, “RA” denotes the rectified accuracy for DivideMix.

Methods	20% noise			50% noise		
	DA	RA	mAP	DA	RA	mAP
ADP-DivideMix	94.2	85.8	56.8	82.8	85.6	52.0
ADP-Clean	–	–	61.6	–	–	56.5
ADP-LCNL	**98.9**	–	**64.9**	**99.7**	–	**59.8**

4.5.2 The Detection Ability of CNL

After revealing the importance of CNL-oriented solutions, we further study the ability of LCNL to detect NA and NC *w.r.t.* different noise ratios. To distinguish the clean samples from the noisy ones, we simply set 0.5 as the confidence boundary for distinguishing the clean from noisy samples. In the evaluation on NC, we report the pair construction statistics and investigate the detection ability of the pair division module. As illustrated in Fig. 5, the detection accuracies on NA and NC keep relatively stable even with the increasing noise ratios, which fully demonstrates the effectiveness of our method.

4.5.3 Fine-grained Ablation Studies

The importance of all the LCNL modules has been well investigated in Table 4, 6 and 8. For a more comprehensive study, we conduct more ablation studies at a finer-grained level.

Effect of the Co-modeling Module: To investigate the effect of the co-modeling module, we replace the co-modeling module by adopting only one network (*i.e.*, self-training manner) or using the teacher-student architecture (*i.e.*, EMA manner (Tarvainen and Valpola, 2017)). As shown in Table 10, both the self-training and EMA manners probably accumulate the bias during NA modeling, thus degrading the performance.

Effect of the Recast Functions: To handle the homogeneous pair combination, we design different kinds of recast functions (Eq. 16). Here, we investigate their roles in Table 11. One could see that the "LCNL-Weighty" setting achieves superior performance thanks to the favorable optimization properties on handling hard-sampling triplets, *i.e.*, Theorem 1and 2.

Effect of the Adaptive Quadruplet Loss: After pair division and correspondence rectification, the robustness of NC is guaranteed by the adaptive quadruplet loss (Eq. 9) which consists of two loss components. Here, we investigate their effects and summarize the results in Table 12. The results demonstrate the importance of the two loss components in achieving robustness.

Table 10 The effects of different training manners.

Manners	20% noise			50% noise		
	R-1	mAP	mINP	R-1	mAP	mINP
LCNL-SelfTraining	65.3	62.4	48.4	62.1	58.7	44.1
LCNL-EMA	66.1	62.9	48.9	58.6	55.8	41.0
LCNL-CoModeling	**67.2**	**64.9**	**51.7**	**62.4**	**59.8**	**45.9**

Table 11 The effects of different recast functions.

Functions	20% noise			50% noise		
	R-1	mAP	mINP	R-1	mAP	mINP
LCNL-Mean (σ_1)	66.3	64.1	50.7	60.3	58.7	45.3
LCNL-Max (σ_2)	**67.4**	64.8	51.6	61.0	58.8	45.0
LCNL-Min (σ_3)	65.2	63.0	49.8	61.3	59.0	45.2
LCNL-MaxMin (σ_4)	65.5	63.3	50.1	61.7	59.5	45.8
LCNL-Weighty (σ_5)	67.2	**64.9**	**51.7**	**62.4**	**59.8**	**45.9**

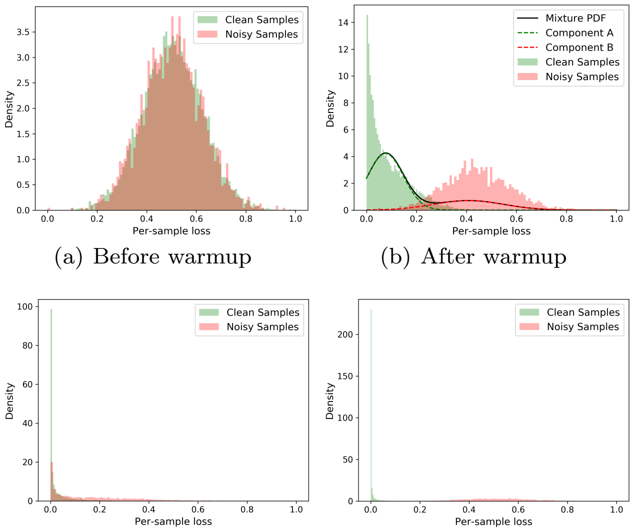
4.5.4 Visualization on Robustness

In this section, we qualitatively study the robustness of LCNL.

Table 12 The effect of the two loss components of the adaptive quadruplet loss.

$\mathcal{L}^{aqdr}$		20% noise			50% noise		
$\mathcal{L}^{atri}$	$\mathcal{L}^{qdt}$	R-1	mAP	mINP	R-1	mAP	mINP
×	×	66.3	63.8	49.8	60.9	57.3	42.3
✓	×	66.7	64.0	50.7	61.7	59.2	45.4
✓	✓	**67.2**	**64.9**	**51.7**	**62.4**	**59.8**	**45.9**

Robustness against NA: We visualize the per-sample identity loss (Eq. 2) on different stages or settings in Fig. 6. From the results, one could see that the vanilla identification loss will overfit the noisy samples (Fig. 6(c)). In contrast, the proposed soft identification loss (Eq. 8) not only fits the clean samples well but also prevents the overfitting on the noisy samples (Fig. 6(d)).



(a) Before warmup

(b) After warmup

(c) Without Eq. 8

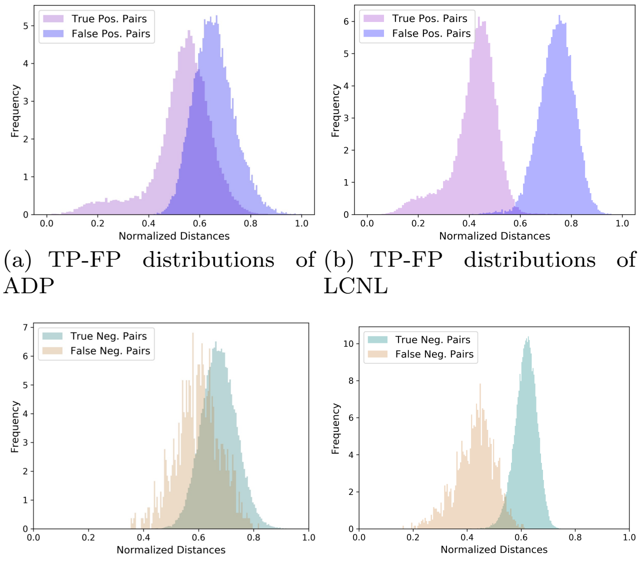
(d) With Eq. 8

Fig. 6 Per-sample loss distribution under different settings.

Robustness against NC: We visualize the pairwise distance distribution of different kinds of pairs *w.r.t.*, ADP and LCNL. As shown in Fig. 7, ADP will confuse the clean and noisy pairs, thus overfitting the NC. In contrast, LCNL not only remarkably distinguishes the pairs but also correctly utilizes the noisy pairs, *i.e.*, increasing the distance of FP pairs while decreasing the one of FN pairs.

4.5.5 Robustness and Generalizability of LCNL

LCNL is a generalized framework, which could endow most existing object ReID methods with robustness



(a) TP-FP distributions of ADP

(b) TP-FP distributions of LCNL

(c) TN-FN distributions of ADP

(d) TN-FN distributions of LCNL

Fig. 7 Pairwise distance distributions of "TP + FP" pairs and "TN + FN" pairs computed through ADP and LCNL.

against the CNL. In this section, we verify such generalizability by additionally developing a robust version of CrossAGW (Ye et al., 2021b) under the LCNL framework which is denoted as "CrossAGW+LCNL". Moreover, we investigate the robustness of "ADP+LCNL" and "CrossAGW+LCNL" with different noise ratios by increasing it from 0% to 60% with an interval of 10%. As illustrated in Fig. 8, LCNL not only endows both CrossAGW and ADP with robustness on the CNL but also performs quite stably under different noise ratios.

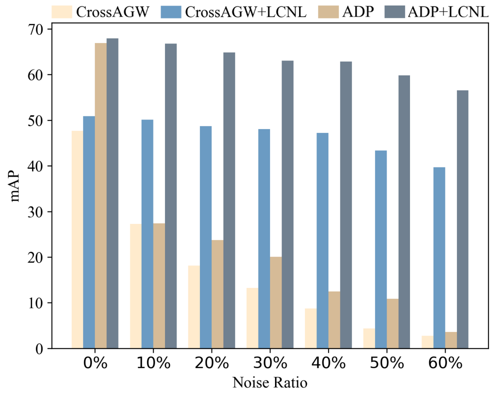


Fig. 8 Performance comparisons of CrossAGW, ADP, AGW+LCNL and ADP+LCNL on the SYSU-MM01 dataset with varying noise ratios.

5 Conclusion

In this paper, we reveal a new problem for object ReID, *i.e.*, coupled noisy labels, which we refer to as noisy annotation and the accompanied noisy correspondence. To tackle this challenge, we propose a CNL-robust framework dubbed learning with coupled noisy labels. The proposed LCNL first estimates the truly-annotated confidences and then rectifies the noisy correspondences. After that, it further groups training pairs into four partitions and achieve CNL-robust object ReID with a provable CNL-robust objective function. Extensive experiments on three different ReID tasks verify the effectiveness of LCNL. As many applications such as sketch-based image retrieval require to annotate samples and construct the training pairs using the annotation, they probably encounter the CNL problem. Therefore, we plan to explore the characteristic of these tasks and study a more general solution for the CNL problem in the future.

Acknowledgements The authors would like to thank the associate editor and reviewers for the constructive comments and valuable suggestions that remarkably improve this study. This work was supported in part by NSFC under Grant U21B2040, 62176171; and in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303.

References

Arpit D, Jastrzebski S, Ballas N, Krueger D, Bengio E, Kanwal MS, Maharaj T, Fischer A, Courville A, Bengio Y, et al. (2017) A closer look at memorization in deep networks. arXiv:170605394 2, 6

Bai S, Bai X, Tian Q (2017) Scalable person re-identification on supervised smoothed manifold. In: CVPR, pp 2530–2539 1

Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel CA (2019) Mixmatch: A holistic approach to semi-supervised learning. NeurIPS 32 4

Chen TS, Liu CT, Wu CW, Chien SY (2020) Orientation-aware vehicle re-identification with semantics-guided part attention network. In: ECCV, Springer, pp 330–346 3

Choi S, Lee S, Kim Y, Kim T, Kim C (2020) Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: CVPR, pp 10257–10266 2, 3, 4, 6

Chu R, Sun Y, Li Y, Liu Z, Zhang C, Wei Y (2019) Vehicle re-identification with viewpoint-aware metric learning. In: ICCV, pp 8282–8291 3

Ge Y, Chen D, Li H (2020) Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. ICLR 1, 2, 3, 4, 6, 7, 8, 12

Goldberger J, Ben-Reuven E (2017) Training deep neural-networks using a noise adaptation layer 4

Han B, Yao Q, Yu X, Niu G, Xu M, Hu W, Tsang I, Sugiyama M (2018) Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: NeurIPS, pp 8527–8537 4, 7

Hao X, Zhao S, Ye M, Shen J (2021) Cross-modality person re-identification via modality confusion and center aggregation. In: ICCV, pp 16403–16412 3

He S, Luo H, Wang P, Wang F, Li H, Jiang W (2021) Transreid: Transformer-based object re-identification. In: ICCV, pp 15013–15022 1, 2, 3, 4, 6, 8, 12

Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv:170307737 8, 10, 12

Hou R, Ma B, Chang H, Gu X, Shan S, Chen X (2019) Interaction-and-aggregation network for person re-identification. In: CVPR, pp 9317–9326 3

Hu P, Peng X, Zhu H, Zhen L, Lin J (2021) Learning cross-modal retrieval with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5403–5413 4

Hu P, Huang Z, Peng D, Wang X, Peng X (2023) Cross-modal retrieval with partially mismatched pairs. IEEE Transactions on Pattern Analysis and Machine Intelligence 5

Huang Z, Niu G, Liu X, Ding W, Xiao X, Peng X (2021) Learning with noisy correspondence for cross-modal matching. In: NeurIPS 5, 6

Kim Y, Yun J, Shon H, Kim J (2021) Joint negative and positive learning for noisy labels. In: CVPR, pp 9442–9451 4

Li H, Wu G, Zheng WS (2021) Combined depth space based architecture search for person re-identification. In: CVPR, pp 6729–6738 3

Li J, Socher R, Hoi SC (2020) Dividemix: Learning with noisy labels as semi-supervised learning. arXiv:200207394 4, 6, 14

Lin Y, Yang M, Yu J, Hu P, Zhang C, Peng X (2023) Graph matching with bi-level noisy correspondence. In: ICCV 5

Liu H, Tian Y, Yang Y, Pang L, Huang T (2016a) Deep relative distance learning: Tell the difference between similar vehicles. In: CVPR, pp 2167–2175 3

Liu X, Liu W, Mei T, Ma (2016b) A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: ECCV, Springer, pp 869–884 3, 10

Liu X, Liu W, Mei T, Ma H (2017) Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia 20(3):645–658 3, 10

Lu Y, Wu Y, Liu B, Zhang T, Li B, Chu Q, Yu N (2020) Cross-modality person re-identification with shared-specific feature transfer. In: CVPR, pp 13379–13389 2, 3, 4, 6

Luo C, Song C, Zhang Z (2022) Learning to adapt across dual discrepancy for cross-domain person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence 1

Ma X, Huang H, Wang Y, Romano S, Erfani S, Bailey J (2020) Normalized loss functions for deep learning with noisy labels. In: ICML, pp 6543–6553 4

Mandal D, Biswas S (2020) Cross-modal retrieval with noisy labels. In: ICIP, IEEE, pp 2326–2330 4

Meng D, Li L, Liu X, Li Y, Yang S, Zha ZJ, Gao X, Wang S, Huang Q (2020) Parsing-based view-aware embedding network for vehicle re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7103–7112 3, 12

Nguyen DT, Hong HG, Kim KW, Park KR (2017) Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors 17(3):605 10

Nguyen DT, Mummadi CK, Ngo TPN, Nguyen THP, Beggel L, Brox T (2019) Self: Learning to filter noisy labels with self-ensembling. In: ICLR 4

Park H, Lee S, Lee J, Ham B (2021) Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In: ICCV, pp 12046–12055 11

Qin Y, Peng D, Peng X, Wang X, Hu P (2022) Deep evidential learning with noisy correspondence for cross-modal retrieval. In: ACM MM 5

Rao Y, Chen G, Lu J, Zhou J (2021) Counterfactual attention learning for fine-grained visual categorization and re-identification. In: ICCV, pp 1025–1034 1, 2, 3, 4, 6, 8

Shen Y, Sanghavi S (2019) Learning with bad training data via iterative trimmed loss minimization. In: ICML, PMLR, pp 5739–5748 4

Shen Y, Xiao T, Li H, Yi S, Wang X (2018) End-to-end deep kronecker-product matching for person re-identification. In: CVPR, pp 6886–6895 3

Shi J, Zhang Y, Yin X, Xie Y, Zhang Z, Fan J, Shi Z, Qu Y (2023) Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In: ICCV 2

Song H, Kim M, Park D, Lee JG (2020) Learning from noisy labels with deep neural networks: A survey. arXiv:200708199 4

Suh Y, Wang J, Tang S, Mei T, Lee KM (2018) Part-aligned bilinear representations for person re-identification. In: ECCV, pp 402–419 3

Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV, pp 480–496 3

Tarvainen A, Valpola H (2017) Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS 14

Tian X, Zhang Z, Lin S, Qu Y, Xie Y, Ma L (2021) Farewell to mutual information: Variational distillation for cross-modal person re-identification. In: CVPR, pp 1522–1531 2

Wang F, Zuo W, Lin L, Zhang D, Zhang L (2016) Joint learning of single-image and cross-image representations for person re-identification. In: CVPR, pp 1288–1296 3

Wang G, Zhang T, Cheng J, Liu S, Yang Y, Hou Z (2019a) Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: ICCV, pp 3623–3632 3

Wang Z, Wang Z, Zheng Y, Chuang YY, Satoh S (2019b) Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: CVPR, pp 618–626 3

Wei Z, Yang X, Wang N, Gao X (2021) Syncretic modality collaborative learning for visible infrared person re-identification. In: ICCV, pp 225–234 3

Wu A, Zheng WS, Yu HX, Gong S, Lai J (2017) Rgb-infrared cross-modality person re-identification. In: ICCV, pp 5380–5389 2, 3, 10

Wu A, Zheng WS, Gong S, Lai J (2020) Rgb-ir person re-identification by cross-modality similarity preservation. International Journal of Computer Vision pp 1765–1785 3

Wu Q, Dai P, Chen J, Lin CW, Wu Y, Huang F, Zhong B, Ji R (2021) Discover cross-modality nuances for visible-infrared person re-identification. In: CVPR, pp 4330–4339 2, 3, 11

Xiao T, Xia T, Yang Y, Huang C, Wang X (2015) Learning from massive noisy labeled data for image classification. In: CVPR, pp 2691–2699 4

Yang M, Li Y, Huang Z, Liu Z, Hu P, Peng X (2021) Partially view-aligned representation learning with noise-robust contrastive loss. In: CVPR 4

Yang M, Huang Z, Peng H, Li T, Lv JC, Peng X (2022a) Learning with twin noisy labels for visible-infrared person re-identification. In: CVPR 4, 5, 11

Yang M, Li Y, Hu P, Bai J, Lv JC, Peng X (2022b) Robust multi-view clustering with incomplete information. IEEE Transactions on Pattern Analysis and Machine Intelligence 4

Ye M, Yuen PC (2020) Purifynet: A robust person re-identification model with noisy labels. IEEE Transactions on Information Forensics and Security 15:2655–2666 2, 3, 4, 7, 12

Ye M, Wang Z, Lan X, Yuen PC (2018) Visible thermal person re-identification via dual-constrained top-ranking. In: IJCAI, vol 1, p 2 3

Ye M, Shen J, Crandall DJ, Shao L, Luo J (2020) Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: ECCV 3, 11

Ye M, Ruan W, Du B, Shou MZ (2021a) Channel augmented joint learning for visible-infrared recognition. In: ICCV, pp 13567–13576 2, 3, 4, 6, 8, 11, 14

Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SC (2021b) Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence 1, 2, 3, 4, 6, 8, 10, 11, 12, 15

Ye M, Li H, Du B, Shen J, Shao L, Hoi SC (2022) Collaborative refining for person re-identification with label noise. IEEE Transactions on Image Processing 31:379–391 2, 3, 4, 7, 12

Yu T, Li D, Yang Y, Hospedales TM, Xiang T (2019) Robust person re-identification by modelling feature uncertainty. In: ICCV, pp 552–561 2, 3, 12

Zhang X, Zhang R, Cao J, Gong D, You M, Shen C (2020) Part-guided attention learning for vehicle instance retrieval. IEEE Transactions on Intelligent Transportation Systems 3, 12

Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015a) Scalable person re-identification: A benchmark. In: ICCV, pp 1116–1124 10

Zheng L, Zhang H, Sun S, Chandraker M, Yang Y, Tian Q (2017a) Person re-identification in the wild. In: CVPR, pp 1367–1376 3

Zheng WS, Gong S, Xiang T (2012) Reidentification by relative distance comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(3):653–668 1

Zheng WS, Gong S, Tao X (2015b) Towards open-world person re-identification by one-shot group-based verification. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(3):591–606 2

Zheng WS, Hong J, Jiao J, Wu A, Zhu X, Gong S, Qin J, Lai J (2022) Joint bilateral-resolution identity modeling for cross-resolution person re-identification. International Journal of Computer Vision pp 136–156 2

Zheng Z, Zheng L, Yang Y (2017b) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV, pp 3754–3762 3, 10, 11

APPENDIX

1 Proof to Theorem 2

Theorem 2 *For the FP&TN combination, the gradient value of $\mathcal{L}^{aqdr}$ with σ_5 w.r.t. d_{ij} is greater than that w.r.t. d_{is} when $d_{ij} < d_{is}$.*

Proof For the FP&TN combination, $\widetilde{y}_{ij}^p = 0$ and $\widetilde{y}_{is}^p = 0$, the gradient of $\mathcal{L}^{aqdr}$ with σ_5 *w.r.t.* d_{ij} is in the form of

$$\frac{\partial \mathcal{L}^{aqdr}}{\partial d_{ij}} = \frac{\exp(2d_{is}) + (1 + d_{is} - d_{ij}) \exp(d_{ij} + d_{is})}{(\exp(d_{ij}) + \exp(d_{is}))^2},$$

and the gradient of $\mathcal{L}^{aqdr}$ with σ_5 *w.r.t.* d_{is} is in the form of

$$\frac{\partial \mathcal{L}^{aqdr}}{\partial d_{is}} = \frac{\exp(2d_{ij}) + (1 + d_{ij} - d_{is}) \exp(d_{ij} + d_{is})}{(\exp(d_{ij}) + \exp(d_{is}))^2}.$$

Let G be the square difference between the values of $\partial \mathcal{L}^{aqdr}/\partial d_{ij}$ and $\partial \mathcal{L}^{aqdr}/\partial d_{is}$, it could be proved that $G > 0, \forall d_{ij} < d_{is}$ by

$$\begin{aligned} G &= \left|\frac{\partial \mathcal{L}^{aqdr}}{\partial d_{ij}}\right|^2 - \left|\frac{\partial \mathcal{L}^{aqdr}}{\partial d_{is}}\right|^2 \\ &= \frac{2(d_{is} - d_{ij}) \exp(d_{ij} + d_{is}) + \exp(2d_{is}) - \exp(2d_{ij})}{(\exp(d_{ij}) + \exp(d_{is}))^2} \\ &> 0. \end{aligned}$$

Therefore, the gradient value of $\partial \mathcal{L}^{aqdr}/\partial d_{ij}$ is greater than $\partial \mathcal{L}^{aqdr}/\partial d_{is}$ when $d_{ij} < d_{is}$.

2 Discussion

Due to the hard mining strategy, the pairs are susceptible to be with noisy correspondence in the presence of noisy annotation, as discussed in Section 3.4.2 in the manuscript. Therefore, the number of different triplet combinations would be inevitably inconsistent. Fortunately, thanks to the proposed adaptive loss $\mathcal{L}^{aqdr}$ (Eq. 9), there is no need to use additional techniques to balance the triplet combinations. Specifically, LCNL adopts loss $\mathcal{L}^{aqdr}$ to adaptively transform the noisy combinations (FP&FN, TP&FN, and FP&TN) into new "clean" combination (TP&TN) for achieving robustness. Thanks to the mechanism of $\mathcal{L}^{aqdr}$, different types of combinations would be transformed into the new "clean" combination (TP&TN), thus having the same importance as each other. As a result, LCNL could achieve robustness against noisy correspondence under imbalanced triplet combinations.

3 More Experiment Details

In the Appendix, we elaborate on the details of the used five datasets as follows.

- SYSU-MM01: It is a large-scale VI-ReID dataset where the images are captured by four visible cameras and two near-infrared ones under both indoor and outdoor environments on the SYSU campus. In the dataset, 22,258 visible images and 11,909 infrared images from 395 identities are used for training, 301 randomly sampled visible gallery images, and 3,803 infrared query images from another 96 identities are used for single-shot evaluation.
- RegDB: It is a VI-ReID dataset that consists of 8,240 images from 412 identities. Each identity has 10 visible and 10 infrared images captured by a dual-camera (one visible and one infrared) system. The standard evaluation protocol is using 10 different training/testing splits. At each evaluation trial, half of the identities are chosen for training and the rest are used for evaluation.
- Market-1501: It is a large-scale V-ReID benchmark, which consists of 32,668 images of 1501 identities captured by six different cameras. In the dataset, 751 identities are used for training and the rest 750 identities are utilized for testing. In the standard testing protocol, 3,368 query images are chosen as the probe set to find the correct matching over 15,913 reference gallery images.
- DukeMTMC: It is a large-scale V-ReID dataset collected from eight different high-resolution videos. This dataset consists of 16,522 training images from 702 identities, 2,228 query images, and 17,661 gallery images from another 702 identities.
- VeRi-776: It is a widely-used dataset for vehicle ReID which is collected in the real-world urban surveillance scenario. The dataset consists of 37,715 training images from 576 identities, 11,579 gallery images, and 1,678 query images from another 200 identities.

4 More Experiment Results

To investigate the impact of different network initialization on our LCNL, we change the initialization difference by varying the hyper-parameters of the default initialization scheme. Accordingly, we conduct experiments with three settings to investigate the impact of initialization differences on the final performance. Specifically, we initialize two networks with 1) the same initialization; 2) different initialization; and 3) relatively different initialization. The results are summarized in

Table 13, where one could find that moderately varying the initialization between two networks might benefit the co-modeling scheme thus slightly improving the performance. However, over-changing the hyper-parameters of the default initialization scheme might lead to unstable optimization. Therefore, in the main experiments, we still initialize networks with default hyper-parameters.

Table 13 Ablation studies the network initialization scheme under SYSU-MM01 with 20% noise. The default setting is marked in gray. The ↓ and ↑ denote the performance degradation and improvement compared to the default setting, respectively.

Initialization Variants			20% noise	
Same	Different	Relatively Different	mAP	mINP
	✓		64.9	51.7
✓			64.5(↓ 0.4)	51.2(↓ 0.5)
		✓	65.3(↑ 0.4)	52.3(↑ 0.6)