Graph-Based Self-Learning for Robust Person Re-identification

Yuqiao Xian[1*] , Jinrui Yang[2] , Fufu Yu[2] , Jun Zhang[2] , and Xing Sun[2†]

[1] School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Youtu Lab, Tencent
xianuyq3@mail2.sysu.edu.cn, {jinruiyang,fufuyu,bobbyjzhang,winfredsun}@tencent.com

Abstract

Existing deep learning approaches for person re-identification (Re-ID) mostly rely on large-scale and well-annotated training data. However, human-annotated labels are prone to label noise in real-world applications. Previous person Re-ID works mainly focus on random label noise, which doesn't properly reflect the characteristic of label noise in practical human-annotated process. In this work, we find the visual ambiguity noise is more common and reasonable noise assumption in annotation of person Re-ID. To handle the kind of noise, we propose a simple and effective robust person Re-ID framework, namely Graph-Based Self-Learning (GBSL), to iteratively learn discriminative representation and rectify noisy labels with limited annotated samples for each identity. Meanwhile, considering the practical annotation process in person Re-ID, we further extend the visual ambiguity noise assumption and propose a type of more practical label noise in person Re-ID, namely the tracklet-level label noise (TLN). Without modifying network architecture or loss function, our approach significantly improves the robustness against label noise of the Re-ID system. Our model obtains competitive performance with training data corrupted by various types of label noise and outperforms the existing methods for robust Re-ID on public benchmarks.

1. Introduction

Person Re-ID [5, 40, 39, 38] is a fine-grained retrieval task that aims to match people across non-overlapping camera views. Impressive progress on the Re-ID task has been made recently with the development of deep convolutional neural networks (deep CNNs) [43, 4]. However, their successes highly rely on high-quality supervision of cleanly labeled data. In real-world industrial applications, label noise is pervasive due to the limited expertise of human annotators and the ambiguity of pedestrian's appearance, leading to obvious performance degradation of existing supervised methods for Re-ID. Designing Re-ID systems tolerant of label noise can help us avoid the labor-intensive and time-consuming manual data cleaning.

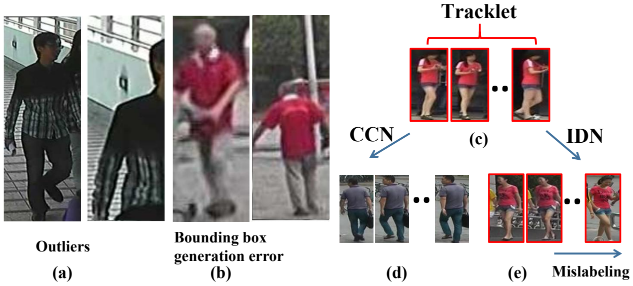


Figure 1. Illustration of different noise in Re-ID datasets. (a) and (b) are mainly caused by person detector. The arrow represents how the noise is generated in the annotation. Person (c) mislablled to person (d) is an example of class-conditional noise (CCN). Person (c) mislablled to person (e) is an example of instance-dependent noise (IDN). (c), (d) and (e) are different person.

Noise problem can be divided into two categories in person Re-ID. The first type is sample noise. The person images are often cropped by off-the-shelf person detectors in current person Re-ID datasets. Due to the effect of person detectors, as shown in Fig. 1 (a) and (b), it may generate some outliers or imperfect person bounding box. Fortunately, with the rapid development of person detection algorithms [46, 13, 18], such sample noise can be readily detected and rectified.

The second noise is label noise, which means that the person image may be incorrectly labeled as another identity. Compared to sample noise, label noise can cause obvious performance degradation to person Re-ID model. More specifically, the common label noise has two types: *class-conditional noise* (CCN) [44, 28, 27] and *instance-dependent noise* (IDN) [3, 6, 7]. CCN is assumed that the noise is independent of image features given the true label. As shown in Fig. 1, that is to say the probability of person (c) being mislabeled as person (d) and person (e) is

[*]This work was done when Yuqiao was an intern at Tencent Youtu Lab.
[†]Corresponding author.

equal under CCN assumption. Previous works(e.g., [42, 41]) mainly consider the CCN in person Re-ID. But we find CCN is a very small proportion in practical human annotation. It is easy to know the possibility of person (c) being wrongly annotated as person (d) that is relatively low. This is because the person (c) and (d) have obviously different visual appearances. On the other hand, we can find that person (c) is incorrectly labeled as person (e) is common in human annotation due to their similar visual appearances. Therefore we argue the IDN is a main label noise in practical person Re-ID scenario.

What's more, the IDN mainly occurs at the image level in image classification problems [3]. Compared to image classification, IDN may bring a more severe negative effect in person Re-ID, which may cause a sequence of image mislabelling. In practice, training images for Re-ID are pedestrian bounding boxes detected and sampled from successive frames in a surveillance video, which is called the *tracklet* [23] of a pedestrian. Human annotators are asked to match the identity of cross-view tracklets captured by non-overlapping cameras. Label noise in Re-ID is more likely to be tracklet-level instead of image-level. That is if a human annotator fails to recognize a person in a tracklet, images in the whole tracklet (video clip) will be assigned an incorrect label (as shown in Fig. 1 (c) , (d) and (e)). Therefore, only considering the the IDN at image-level may not reflect the characteristics of label noise in the process of real human labeling. To address this issue, we further extend the IDN assumption, and propose a new type of label noise, namely the tracklet-level label noise (TLN) , which can provide a more realistic description about the label noise of human annotation in real-world person Re-ID system.

Although a variety of methods have been developed for robust deep learning with noisy labels, most of them focus on image classification [21, 11, 1, 31]. Two underlying assumptions limit their application to the Re-ID problem: 1) They assume that human annotation errors happen in image-level, which may not conform to the annotation process of Re-ID. 2) They assume that there are enough training samples for each class. Conversely, Re-ID is a few-shot problem that usually has more identities (IDs) and much fewer samples for each class (i.e., identity) , as shown in Table 1.

Therefore, based on the above analysis, to address the label noise problem in person Re-ID with limited samples for each identity, we propose a simple and effective framework, namely Graph-Based Self-Learning (GBSL), to iteratively detect and rectify false annotations in the deep representation learning process. We build a relational graph based on nearest neighbors and propagate the label messages to rectify inconsistent labels in the iteration of model training. After label correction, the network is provided with labels of higher quality which facilitates learning more discriminative features for label correction in the next iteration.

Table 1. Comparison on the number of classes and labeled training images in general image classification (left) and person Re-ID (right).

Classification	#classes	#imgs	Re-ID	#IDs	#imgs
CIFAR-10	10	5K	Market [47]	751	17.2
CIFAR-100	100	0.5K	Duke [48]	702	23.5
Clothing1M	14	71.4K	MSMT [36]	1041	31.3
Food-101N	101	750	Real-world	Massive	Few

To summarize, our main contributions are: 1) We discuss a more practical label noise in real-world person Re-ID, **t**racklet-**l**evel label **n**oise (TLN), for the first time. TLN is a type of label noise in the tracklet-wise annotation process of person Re-ID. 2) Relaxing the constraints of noise rate or auxiliary clean data, a model-agnostic self-learning framework is presented to automatically correct various types of noisy labels, which can be embedded into most person Re-ID models easily. 3) On public Re-ID benchmarks corrupted by different types of severe label noise, the proposed method surpasses all the compared methods for robust person Re-ID by a clear margin.

1.1. Related Work

This work is closely related to noise-robust person Re-ID, tracklet person Re-ID, and robust deep learning with noisy labels.

Noise-robust Re-ID. Developing noise-robust Re-ID is a critical issue because open-world Re-ID applications usually suffer from unavoidable noise in data collection and human annotation [40], including sample noise and label noise. Sample noise is mainly caused by inaccurate detection or tracking algorithms, includes outlying regions (*e.g.* occlusion and background) within the bounding box [30] and outlier frames within each tracklet [2]. Attention mechanism and pose-guided methods are posed to handle noise within an image. For outliers frames in the video sequence, frame re-weighting and spatial-temporal attention are studied in recent research. Label noise in Re-ID has also been investigated in some preliminary works (namely *robust person Re-ID*) [42, 41]. They only consider image-level random noise (similar to CCN). Methodologically, Yu *et. al.* [42] focus on robust architecture design, while Ye and Yuen [41] focus on label refinement and sample re-weighting based on the model prediction. In comparison, beyond the image-level CCN, we start the first attempt to study a more realistic and challenging tracklet-level noise model and propose a label correction method based on graph consistency.

Tracklet Person Re-ID The tracklet association of pedestrian images can be useful supervision for unlabeled dataset [22, 23, 37]. The noisy frames within a tracklet, which is a type of sample noise produced by detection algorithms, are also studied in [24]. Instead, we focus on the label noise in supervised person Re-ID with TLN generated from the false human annotation in this paper.

Robust Deep Learning with Label Noise Training deep neural networks with noisy labels has been widely explored in recent years [1, 31], including robust losses [44, 27, 25], robust model architectures [9, 10], sample re-weighting [45, 17], label correction and others. Most of these works focus on robust deep learning for image classification, which require a set of clean labels [15, 21, 35] and relying on noise distribution assumptions. Specific design on network architectures or loss functions also limits the applications to other vision tasks. From the perspective of noise distribution, many previous methods only consider random or class-conditional noise, while recent researches [3, 6] point out that label noise pattern in the real-world is most likely to be instance-dependent. Methodologically, there exist iterative self-learning frameworks for noisy label learning which embeds re-weighting [41], filtering [11, 34] or label correction [12] in the representation learning process. In comparison, we propose a graph-based label correction method that employs message propagation to rectify noisy labels.

1.2. Preliminary

In this section, to formulate the problem of person Re-ID with noisy labels, we revisit the commonly studied CCN and IDN. Based the IDN assumption, we further propose a more realistic and challenging TLN in person Re-ID. Notably, we assume that the label noise rate ρ_{noise} is unknown and no auxiliary clean data is available.

1.3. Class-Conditional Label Noise (CCN) and Instance-Dependent Label Noise (IDN)

Traditionally, the noise transition matrix $T(X)$ is introduced to model the distribution of noisy labels. X denotes the sample feature. The transition matrix of CCN is formulated as,

$$T_{i,j}(X) = \mathbb{P}(\tilde{Y} = j | Y^* = i), \tag{1}$$

where the labels of samples is flipped to noisy labels $\tilde{Y}$ with a probability only depending on their ground truth Y^*.

The IDN describes label flipping that depends on the inherent input features, whose transition matrix can be formulated as,

$$T_{i,j}(X) = \mathbb{P}(\tilde{Y} = j | Y^* = i, X), \tag{2}$$

which is a function of both Y^* and X. When Y^* is given, the transition matrix only depends on X. Thus we can intuitively know the IDN is closely related to visual ambiguity problem in person Re-ID.

The transition matrix models the image-level label flipping probability. In a general image classification task, we usually have no prior knowledge of the association among different samples. Human annotators usually label the images one by one. Therefore generating the image-level label noise for image classification is reasonable. However, both CCN and IDN can not well describe the tracklet-wise label noise in human annotation process of person Re-ID.

1.4. Tracklet-Level Label Noise (TLN)

Given a tracklet $\mathbf{S} = \{x_1, x_2, ...\}$, which is a set of images sampled from a sequence of bounding boxes detected from consecutive frames of surveillance video, the human annotators match it with another tracklet captured from other cameras and assign an identical identity label y to all the images.

Definition 1 (TLN Model). *If a bounding box image of a person is labeled with an incorrect label j, all images within the same tracklet will be assigned the same label j, i.e., $\exists x_i \in \mathbf{S}', \tilde{y}_i = j \neq y_i^* \iff \forall x_k \in \mathbf{S}', \tilde{y}_k = j$.*

The TLN model formulates a constraint to the generation of label flipping, *i.e.*, images within a tracket should have coherent annotations. Since images within a tracklet are usually of high similarity, the distribution of TLN is *locally-concentrated.*

To model realistic TLN in a dataset, we first pre-train a model on the clean dataset and use it to find the most similar identity except the ground truth identity (namely the secondary identity) of each image based on the outputs of classifier. The process is similar to the IDN [3] generation. Then, the secondary identity of a tracklet is determined by the most frequent secondary identity of all images in the tracklet. When generating TLN, we will change all the labels of images in the same tracklet into the secondary identity of the tracklet. Details about generation of three kinds of label noise are provided in supplementary materials.

2. Methodology

2.1. Iterative Self-Learning Framework

Our goal is to learn discriminative features for person Re-ID with noisy human annotations. Fig. 2 illustrates the proposed graph-based self-learning framework which iteratively optimizes network parameters $\mathbf{\Theta}$ and rectifies the labels of the noisy dataset $\tilde{Y}$. In the network optimization phase, we train a deep network to learn discriminative representations which helps us distinguish clean and noisy labels. In the label correction phase, we construct a similarity graph to detect and rectify inconsistent labels, which can learn better discriminative features. After several iterations between network optimization and label corrections, the labels converge and then we can continue training the network parameters with stable labels until the convergence of the model.

Our method is model-agnostic and loss-independent which focuses on label correction. So we adopt a widely used ResNet-50 [14] architecture to be optimized with a combination of identity loss (*i.e.* cross entropy loss) L_{id} and hard triplet loss [16] L_{tri}, which is formulated as:

$$\mathbf{\Theta} = \arg\min_{\mathbf{\Theta}} L_{id}(X|\mathbf{\Theta}, \overline{Y}) + L_{tri}(X|\mathbf{\Theta}, \overline{Y}). \tag{3}$$

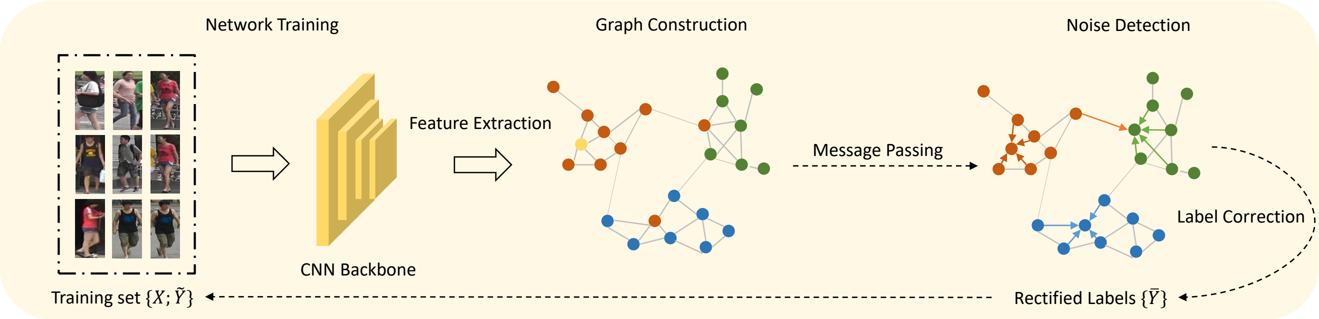


Figure 2. Illustration of the proposed framework Graph-Based Self-Learning (GBSL).

where $\overline{Y}$ is the rectified label which is updated in each iteration of label correction and $\overline{Y} = \tilde{Y}$ in the first iteration.

2.2. Graph-Based Message Passing

Existing label refinement or correction methods for noisy labels rely on model classifier [41, 3] or class prototypes [12] to refine the labels for noisy dataset. However, the training data for Re-ID is usually few-shot and long-tailed, the model classifier will be sensitive to label noise and it's difficult to find reliable prototypes. To address this, we construct a relational graph to propagate label information to find inconsistency in the graph, which may probably detect and rectify the samples with incorrect labels. Based on the *cluster assumption* [51] that nearby points are likely to have the same label, we detect inconsistent points with noisy labels and correct the labels by aggregating messages from their neighbors on the graph.

Graph construction. Given a network with parameters $\mathbf{\Theta}$, we obtain the representation set $Z = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n]$, where $\mathbf{z}_i = \phi(x_i|\mathbf{\Theta})$. We construct a sparse affinity matrix $A \in \mathbb{R}^{n \times n}$ by

$$A_{ij} = \begin{cases} 1 - d(\mathbf{z}_i, \mathbf{z}_j), & \mathbf{z}_j \in N_k(\mathbf{z}_i); \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

where $d(\mathbf{z}_i, \mathbf{z}_j) \in [0, 1]$ is the normalized distance metric (*e.g.*, cosine distance) between two samples band. $N_k(\mathbf{z}_i)$ denotes the set of $\mathbf{z}_i$'s k-nearest neighbors. Generally, an affinity matrix should be symmetric. We can introduce the symmetric affinity matrix A^* with zero diagonal as:

$$A^* = \frac{1}{2}(A + A^\top) \quad (5)$$

Message passing. After constructing the k-nearest neighbor graph, we aim to optimize the label space to be consistent to the contextual information in the feature space. It is performed by label message passing on the k-nearest graph neighbor. We denote the label matrix as L whose row corresponding to the label of each example is one-hot encoded (*i.e.*, $L_{ij} = 1$ if $y_i = j$ otherwise $L_{ij} = 0$). In each label correction iteration, the propagation model for the graph-based message passing is formulated as:

$$L := D^{-\frac{1}{2}}(A^* + \lambda I)D^{-\frac{1}{2}}L \quad (6)$$

where D is the degree matrix of $A^* + \lambda I$. A^*_{ij} measures the connectedness between sample x_i and x_j and controls the weights of message passing between them.

The propagation model is the graph laplacian of the matrix $A^* + \lambda I$. I is the identity matrix where the sample propagates its label to itself. λ is a hyperparameter which controls the degree of *self-reinforcement* in the correction phase. Since most labels are correct in a human-annotated dataset, we should emphasize a sample's own label instead of purely relying on information from its neighbors. Our propagation model is relative to the first-order approximation of spectral graph convolution used in GCN [20]. The difference lies in that we propagate the label message instead of node features. Different from label propagation [52, 51] methods for semi-supervised learning which repeatedly propagate labels from labeled samples to unlabeled samples until converging to a stable state, the label message propagation in our method is embedded into the representation learning of a deep network. We only propagate the label information once in each label correction stage and explicitly perform hard label correction. The principle for this is to inhibit the propagation of noisy information before learning more discriminative representations. Besides, solving our first-order propagation model also needs lower computation costs.

Label correction. After neighborhood aggregation by message passing, if $\arg\max_j L_{ij} \neq y_i$, which means the sample's current label is inconsistent with its neighbors on the graph, we regard the sample is wrongly annotated. Then, we explicitly rectify the label by

$$\overline{y}_i := \arg\max_j L_{ij}, j = 1...C. \quad (7)$$

where C is the number of identity.

In the graph-based label correction, each sample "votes" for the labels of its neighbors according to their similarity. After label correction, the network is provided with labels with higher quality. The rectified labels boost the network to learn more discriminative features which can help to correct more labels in the next iteration. The whole algorithm is summarized in Algorithm 1.

Algorithm 1 Graph-Based Self-Learning (GBSL)

```
Input: Training dataset $X$ with noisy labels $\tilde{Y}$, initialized network parameters $\mathbf{\Theta}$, set of correction epoch $T_C$.
Output: Optimized network parameter $\mathbf{\Theta}$, rectified labels $\overline{Y}$.
for $t = 1; t <= num\_of\_epoch; t++$ do
    if $t \in T_C$ then
        Extract features $Z$ with network encoder.
        Construct (update) relational graph A* by Eq. (4-5)
        Optimize label matrix with first-order message passing by $L := D^{-\frac{1}{2}}(A^* + \lambda I)D^{-\frac{1}{2}}L$ (Eq. (6))
        Detect and rectify inconsistent labels with Eq. (7)
    end if
    Optimize $\mathbf{\Theta}$ with $\overline{y}_1, \overline{y}_2, ..., \overline{y}_n$ (Eq. (3))
end for
return $\overline{Y}$, $\mathbf{\Theta}$.
```

3. Experiments

3.1. Datasets and Evaluation Protocols

Benchmark datasets. To follow with the previous works on robust person Re-ID [42, 41] and analyze different types of label noise, we evaluate our method on two large-scale benchmark datasets for person Re-ID: Market-1501 and DukeMTMC-reID. Market-1501 [47] has 32,688 labeled person images of 1,501 identities collected from 6 different cameras. DukeMTMC-reID [48, 29] contains 36,411 labeled images of 1,404 people from 8 camera views. The images in the two image-based datasets contain the information of camera ids and tracklet from their image names. Taking Market-1501 for example, in image name "0001_c1s1_001051_00.jpg", "0001" is identity. "c1" is the camera id. "s1" is sequence(tracklet) id. In TLN generation, the images of Market-1501 are divided into 3,262 tracklets (and 2,195 tracklets for DukeMTMC-reID).

Evaluation metric. We report the results of the rank-1 accuracy (R1) and mean average precision (mAP) following the standard protocols in [47, 48] without post-processing technique, like re-ranking [49] or multiple query retrieval [47]. We also evaluate the quality of label correction by precision (Pre.) and recall (Rec.). We classify the correction operation into three types: true correction, false correction, and switch correction, as shown in Fig. 3. True correction (TC) means a noisy (incorrect) label is corrected by the algorithm. False correction (FC) means a clean label is changed to a wrong label. Switch correction (SC) means a noisy label is modified to another incorrect label. Then the precision and recall rate of label correction is defined as:

$$Precision = \frac{TC}{TC + FC + SC} \times 100\% \tag{8}$$

$$Recall = \frac{TC - FC}{\rho_{noise} \times |\mathbf{I}|} \times 100\% \tag{9}$$

where $|\mathbf{I}|$ is the size of image dataset.

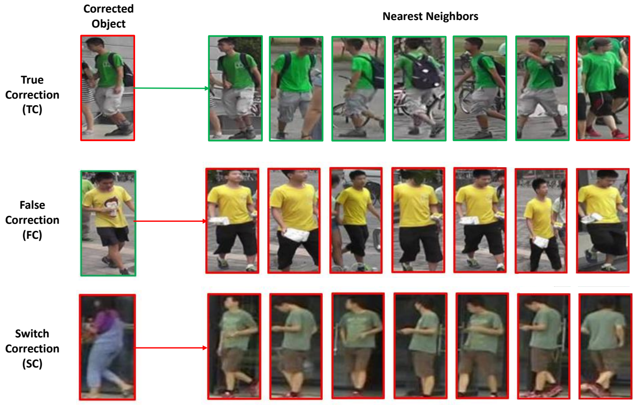


Figure 3. Examples of true correction (TC), false correction (FC) and switch correction (SC) by graph-based message passing. The green box denotes a true label with the object image while the red box denotes a false label with the object.

3.2. Implementation Details

We adopt ImageNet [8] pre-trained ResNet-50 [14] as the backbone of feature encoder, and a linear classifier with BNNeck [26] is added in the last of the network. All images are resized to 256 × 128 with random flipping and random erasing [50] for argumentation. The stride of the last stage in backbone to 1. Adam optimizer [19] is adopted with the batch size of 64 and an initial learning rate of 3.5×10^{-4}, decreasing with a factor of 0.1 in 40^{th} and 70^{th} epoch of 80 epochs in total. We perform label correction every 2 epoch in the first 40 epochs until the labels remain stable after 40^{th} epoch (*i.e.*, $T_C = \{2, 4, ..., 40\}$ in Algorithm 1). The k-reciprocal encoding [49] is adopted as the distance metric in Eq. (5). We implement our experiments with Pytorch 1.6 on a regular PC with a Tesla P40 GPU. The label correction process is implemented on GPU and needs about 30 minutes in the whole training process. After a simple grid search, we set $k = 8$ and $\lambda = 2$ in all the experiments unless otherwise specified. For a fair comparison, the generated label noise for models of all competing methods is fixed.

3.3. Comparison With State-of-the-Arts

We compare our method with two existing methods for robust Re-ID (PurifyNet [41] and DistributionNet [42]) and four popular methods (MeanTeacher [32], Co-Teaching [11], DSL [12] and SEAL [3]) for robust deep learning on Re-ID benchmarks. For a fair comparison, we implement these methods and report the results with the same backbone (the strong baseline for re-ID [26]), except for the robust architecture method DistributionNet. The "noise-free" model is training the baseline model with the original labels of the benchmarks, whose performance is considered as the *upper bound* in our settings. We evaluate the methods of learning with three different types of label noise in Re-ID, including uniformly distributed CCN, IDN, and the proposed TLN. The results are shown in Table 2 - Table 4.

Impact of different types of label noise. By comparing the baseline model with the noise-free model in Table 2 - 4, we have the following observations: (1) The baseline method suffers from an obvious decline of performance with all kinds of label noise. (2) CCN noise is more destructive to the model performance than the same proportion of IDN or TLN. The reason is that CCN has stronger randomness that assigns a random label to a noisy sample which may cause severe feature distortion. (3) An equal proportion of IDN and TLN has a similar impact on the baseline. They have the common operation in assigning the same label to images of a different person with a similar appearance.

Robustness against image-level label noise. The results shown in Table 2-4 demonstrate that our GBSL achieves the best performance among all the compared methods in all types of noisy label settings. In the setting of learning with image-level label noise (CCN and IDN), our method surpasses the competitors by a clear margin and has high precision and recall in label correction (as shown in Fig. 5). We also observe that the image-level instance-dependent label noise in Re-ID is not evidently more difficult than the random CCN for our method, which leads to a different conclusion from that in general image classification with instance-dependent noisy labels. The main reason for that is the person images within a tracklet are usually have high visual similarity, *i.e.*, lies in a narrow region of feature space. If only a small part of image samples in a tracklet are accidentally assigned wrong labels, they can be readily detected and rectify by aggregating label information from their neighbors in GBSL. Neighborhood relationship is not exploited in all compared methods except DSL, resulting in their inferior performance in the task of Re-ID with label noise.

Analysis on tracklet-level label noise. Although TLN is less destructive to the baseline method, it is also prone to be overfitted by the models and it is more difficult to handle through robust deep learning methods. As shown in Table 4, all the methods have a small improvement on the baseline model. Our method outperforms all the competitors on both benchmarks, but still obtains lower performance than that in the setting with the same proportion of CCN or IDN. The intermediate representations learned by the models are not view-invariant, which are only effective in rectifying noisy samples within tracklets.

Table 2. Comparison with other methods on noisily-supervised learning of person Re-ID benchmarks with uniform **class-conditional (random) label noise (CCN)**.

	Market-1501				DukeMTMC			
Method	10% noise		20% noise		10% noise		20% noise	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Noise-free	94.1	86.7	94.1	86.7	86.3	75.9	86.3	75.9
Baseline	87.7	72.8	78.1	58.2	77.3	63.2	65.8	51.2
Dist. Net [42]	82.3	61.5	77.0	53.4	68.6	48.0	62.4	40.9
PurifyNet [41]	85.2	66.2	84.1	64.8	76.5	61.3	74.5	56.2
Co-Teach [11]	84.5	65.3	83.2	63.8	74.2	57.1	62.5	43.8
MeanTeach [32]	87.0	72.3	77.0	57.5	76.0	62.1	64.0	49.3
SEAL [3]	90.2	79.1	84.6	68.7	80.1	66.2	78.2	65.8
DSL [12]	91.2	80.3	89.7	78.6	82.1	71.3	81.5	72.0
Ours	**93.7**	**84.8**	**92.2**	**82.2**	**85.9**	**74.5**	**85.2**	**73.9**

Table 3. Comparison with other methods on noisily-supervised learning of person Re-ID benchmarks with **instance-dependent (patterned) label noise (IDN)**.

	Market-1501				DukeMTMC			
Method	10% noise		20% noise		10% noise		20% noise	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Noise-free	94.1	86.7	94.1	86.7	86.3	75.9	86.3	75.9
Baseline	89.6	76.7	84.1	67.1	79.6	66.7	71.7	57.3
Dist. Net [42]	52.4	27.0	49.3	24.4	37.7	20.8	34.5	18.5
PurifyNet [41]	86.7	70.2	85.3	66.5	77.9	65.1	75.6	60.8
Co-teach [11]	85.2	67.0	84.2	65.3	74.8	58.3	68.3	53.0
MeanTeach [32]	88.7	75.3	83.2	64.6	78.7	65.9	69.9	55.6
SEAL [3]	90.5	78.9	86.6	71.3	81.2	69.1	79.6	67.4
DSL [12]	91.5	81.0	90.2	79.6	84.0	73.0	83.5	72.7
Ours	**93.6**	**84.8**	**91.9**	**82.3**	**86.2**	**75.4**	**85.5**	**74.1**

Table 4. Comparison with other methods on noisily-supervised learning of person Re-ID benchmarks with the proposed **tracklet-level label noise (TLN)**.

	Market-1501				DukeMTMC			
Method	10% noise		20% noise		10% noise		20% noise	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Noise-free	94.1	86.7	94.1	86.7	86.3	75.9	86.3	75.9
Baseline	90.4	78.8	85.3	69.9	81.6	68.9	74.5	60.7
PurifyNet [41]	87.2	71.8	86.5	69.2	78.1	66.1	74.2	59.8
Co-teach [11]	86.3	68.8	83.3	64.7	75.1	60.2	71.2	57.6
MeanTeach [32]	89.8	76.8	84.3	66.8	80.2	68.3	73.4	58.9
SEAL [3]	89.4	77.0	85.4	70.3	81.5	68.3	74.0	59.2
DSL [12]	90.5	79.8	86.1	71.5	81.9	69.9	75.3	62.3
Ours	**92.0**	**81.7**	**88.8**	**76.6**	**82.3**	**70.8**	**76.5**	**65.6**

Evaluation on label correction. Fig. 4 illustrates the label correction during training iterations and Fig. 5 shows the results of GBSL's precision and recall on label correction with comparison to naive *k*-NN classifier and DSL. We can observe that our method can effectively correct the labels of both CCN and IDN, where IDN is only slightly more difficult to detect than CCN even the generation is quite different. The correction precision of CCN and ICN is approximately 90% with a recall over 80%, which means our label correction method can significantly improve the label quality of the corrupted dataset. We can also see that only 17% corrupted labels are successfully rectified by the GBSL and the results of the other two methods are even lower, showing that such type of label noise is much more difficult to detect by the models than the image-level label noise (*i.e.*, CCN and IDN.).

Table 5. Ablation study of propagation model on Market-1501 with both types of label noise.

Method	Propagation Model	20% IDN noise				20% TLN noise			
		ReID		Label Correction		ReID		Label Correction	
		R1	mAP	Pre.	Rec.	R1	mAP	Pre	Rec.
Baseline	-	84.1	67.1	-	-	85.3	69.9	-	-
Model prediction	$\arg\max_j f(x\|\mathbf{\Theta})_j$	85.1	68.4	32.5	22.1	85.6	71.4	25.6	5.4
k-NN classifier	$\arg\max_y \sum_{i=1}^{C} 1[y_i = y, x_i \in N_k(x)]$	87.2	73.9	36.5	26.6	84.9	67.0	23.0	6.9
Label Spread [51]	$(\mathbf{I} - \alpha\mathbf{A})^{-1}\mathbf{L}$	90.5	79.5	68.5	71.2	86.4	71.2	43.2	11.2
Ours w/o $\lambda\mathbf{I}$	$\mathbf{D}^{-\frac{1}{2}}\mathbf{A}^*\mathbf{D}^{-\frac{1}{2}}\mathbf{L}$	88.1	75.8	40.2	28.9	85.6	71.0	25.1	7.8
Ours w/o symmetric $\mathbf{A}^*$	$\mathbf{D}_l^{-\frac{1}{2}}(\mathbf{A} + \lambda\mathbf{I})\mathbf{D}_r^{-\frac{1}{2}}\mathbf{L}$	90.8	79.7	70.2	74.6	86.9	73.2	47.8	16.3
Ours	$\mathbf{D}^{-\frac{1}{2}}(\mathbf{A}^* + \lambda\mathbf{I})\mathbf{D}^{-\frac{1}{2}}\mathbf{L}$	**91.9**	**82.3**	**89.7**	**84.2**	**88.8**	**76.6**	**56.3**	**17.0**

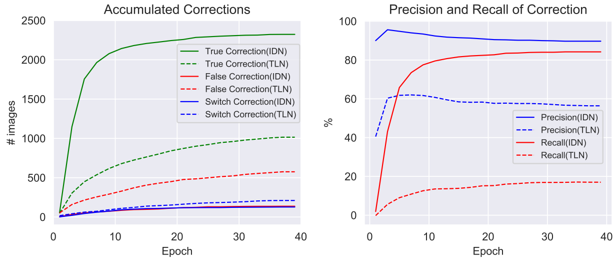


Figure 4. Accumulated corrections (left), correction precision and recall (right) on Market-1501 with 20% IDN noise and TLN noise during training epoch by the proposed method.

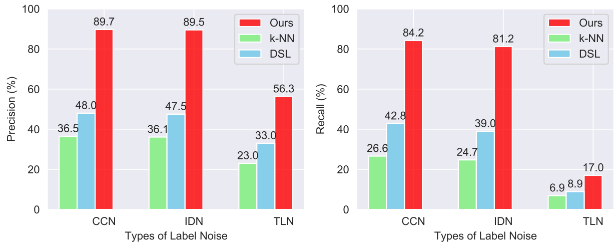
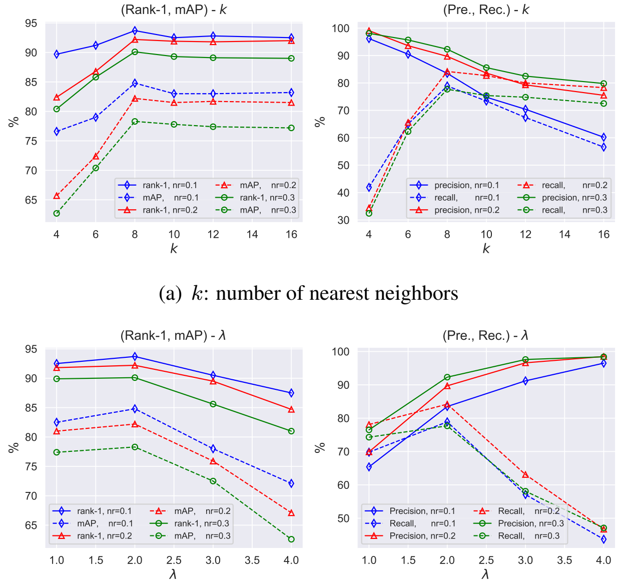


Figure 5. Evaluation of label correction precision (Pre.) and recall (Rec.) on Market-1501 with 20% different types of label noise.

3.4. Component Evaluation and Discussion

Propagation model. We conduct an ablation study using different propagation models in the graph-based self-learning framework, which can be regarded as variants of the proposed GBSL model. The results are shown in Table 5. We have the following observations: (1) Relying on model classifier prediction or naive k-NN classifier to predict labels without considering the similarities among samples, which is easily affected by the noisy labels and renders the spread of noisy label, and causes a low correction precision. (2) Our method surpasses the diffusion model of label propagation for semi-supervised learning, which iteratively passes label information from labeled data to unlabeled data. The reason is that propagating labels message repeatedly within a collection iteration also propagates more noisy labels, resulting in lower correction accuracy. (3) Both self-reinforcement and symmetric affinity matrix improve the performance of label correction.



(a) k: number of nearest neighbors

(b) λ: degree of self-reinforcemen

Figure 6. Hyperparameter analysis with different noise rates (nr).

Sensitivity to k. In Fig. 6 (a), we vary k, *i.e.*, the number of nearest neighbors, from 4 to 16 under different noise ratio (CCN) from 10% to 30%. It should be noticed that k should not be too small otherwise the correction is easily affected by local noisy samples. We observe that the performance is robust when $k \geq 8$. We also observe that the optimal choice for k is 8 and independent from the noise rate.

Sensitivity to λ. In Fig. 6 (b), we analyze another important hyperparameter in our method, the degree of self-reinforcement λ. We find that $\lambda = 2$ is best for all noise rates on both datasets. Using a small λ producing more false correction, resulting in low precision in label correction and the recall is also low because lots of new noisy labels are produced by false correction. Using a large λ will have high precision because we only correct the labels when we are highly confident but will have a low recall rate of noisy correction. Similar to k, the optimal value of λ is also independent of the noise rate.

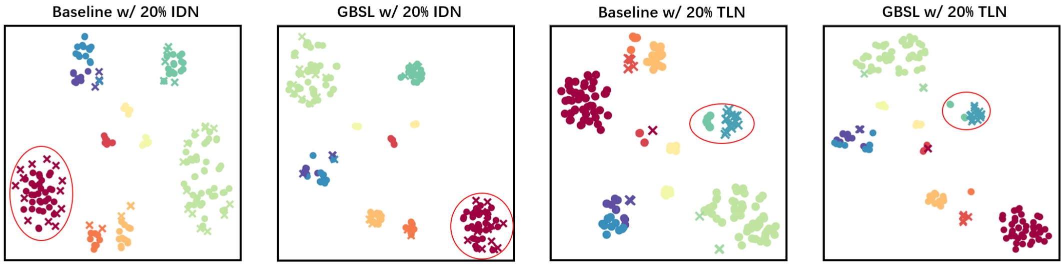


Figure 7. t-SNE visualization of IDN and TLN. We use different colors to denote different identities and crosses to denotes samples with incorrect labels.

Table 6. Evaluation of label correction precision (Pre.) and recall (Rec.) on Market-1501 with 20% different types of label noise.

Method	CCN		IDN		TLN	
	rank-1	mAP	rank-1	mAP	rank-1	mAP
Soft re-labeling	85.3	68.2	86.2	70.5	85.2	68.8
Hard (Ours)	92.2	82.2	91.9	82.3	88.8	76.6

Hard correction vs. soft re-labeling. We validate the effect by training the network with softmax logits of the label matrix after message passing (*i.e.*, $L_{id} = -\frac{1}{n}\sum_{i=1}^{n}\overline{p}_{ij}\log f(x_i|\mathbf{\Theta})$, $\overline{p}_{ij} = \frac{exp(L_{ij}/\tau)}{\sum_j exp(L_{ij}/\tau)}, j = 1...C$). The comparison is in Table 6. We find that using a soft re-labeling obtain inferior performance than hard label correction. We believe the reason is that using a soft re-labeling may make the model easier overfit to the label noise and lose useful information of hard samples. Using the hard correction in our first-order propagation model can cut off the transmission of noisy labels.

Visualization. We randomly choose 10 persons from Market-1501 and visualize their features with t-SNE [33] in Fig. 7. We have the following observations: (1) The baseline method overfits both IDN and TLN with different patterns. Samples with IDN distribute in outlying regions of other samples with the same identity, while samples with TLN prefer to gather in separate regions from other samples with the same identity. (2) Our GBSL model can produce more compact feature clusters than the baseline model, indicating that the proposed method can facilitate robustness against label noise.

Robustness test on clean dataset. We also evaluate our method with the original labels of Market-1501 and DukeMTMC-reID benchmarks, which are relatively clean with limited annotation errors. We observe that the performance is stable compared with the baseline model. On DukeMTMC-reID, the performance is slightly improved, indicating that its training set may be originally noisy. Although GBSL may wrongly modify some of the clean labels in primitive iterations of training, most of them will be later corrected during the self-learning process as the learning of more discriminative features. The rest samples that have been modified but not be corrected lately are mostly outliers of each identity. Their labels will change to an identity of a pedestrian who has a similar appearance to them and such errors may not cause significant harm to the performance. Notably, the proposed method can detect some originally incorrect labels in the benchmark datasets with the proposed algorithm, as shown in Fig. 8. For example, an image of a man wearing a black sling bag, white breast piece, and white shoes is assigned to ID 0939 of Market-1501 who wears black shoes with a similar appearance.

Table 7. Robust test on clean training set of Re-ID benchmarks.

Method	Market-1501		DukeMTMC-reID	
	rank-1	mAP	rank-1	mAP
Baseline	**94.1**	**86.7**	86.3	75.9
Ours	94.0	85.9	**86.7**	**76.8**

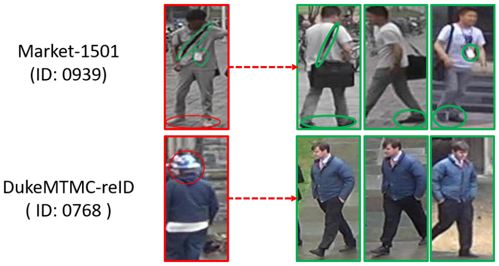


Figure 8. Examples with incorrect annotations in the existing Re-ID benchmarks detected by our method.

4. Conclusion

In this paper, we study the problem of robust person Re-ID with noisy labels. Based on characteristics of the annotation process in person Re-ID, we propose a type of more realistic and challenging noise, TLN. To handle label noise with limited training samples for each identity, we propose a graph-based self-learning framework for robust person Re-ID to iteratively learn discriminative representation and correct inconsistent labels. The proposed method can effectively reduce the IDN and TLN for robust person Re-ID and significantly improve the robustness of a baseline model against label noise. Although our method can well address the image-level label noise in Re-ID, the proposed TLN remains challenging and deserves further investigation in the future.

References

[1] Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771, 2021.

[2] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1169–1178, 2018.

[3] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11442–11450, 2021.

[4] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3300–3310, 2020.

[5] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):392–408, 2017.

[6] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*, 2020.

[7] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *International Conference on Machine Learning*, pages 1789–1799. PMLR, 2020.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2016.

[10] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. *Advances in neural information processing systems*, 31, 2018.

[11] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

[12] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5138–5147, 2019.

[13] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11328–11337, 2021.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31, 2018.

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[17] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Ondřej Chum, and Cordelia Schmid. Graph convolutional networks for learning with few clean and many noisy labels. In *European Conference on Computer Vision*, pages 286–302. Springer, 2020.

[18] Abdul Hannan Khan, Mohsin Munir, Ludger van Elst, and Andreas Dengel. F2dnet: Fast focal detection network for pedestrian detection. *arXiv preprint arXiv:2203.02331*, 2022.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[21] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5456, 2018.

[22] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision (ECCV)*, pages 737–753, 2018.

[23] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1770–1782, 2019.

[24] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised noisy tracklet person re-identification. *arXiv preprint arXiv:2101.06391*, 2021.

[25] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International conference on machine learning*, pages 6226–6236. PMLR, 2020.

[26] 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.

[27] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pages 6543–6553. PMLR, 2020.

[28] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2019.

[29] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for

multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016.

[30] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1179–1188, 2018.

[31] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[32] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[33] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[34] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8688–8696, 2018.

[35] Zhen Wang, Guosheng Hu, and Qinghua Hu. Training noise-robust deep neural networks via meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4524–4533, 2020.

[36] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.

[37] Qiaokang Xie, Wengang Zhou, Guo-Jun Qi, Qi Tian, and Houqiang Li. Progressive unsupervised person re-identification by tracklet association with spatio-temporal regularization. *IEEE Transactions on Multimedia*, 23:597–610, 2020.

[38] Jinrui Yang, Jiawei Zhang, Fufu Yu, Xinyang Jiang, Mengdan Zhang, Xing Sun, Ying-Cong Chen, and Wei-Shi Zheng. Learning to know where to see: a visibility-aware approach for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11885–11894, 2021.

[39] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3289–3299, 2020.

[40] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[41] Mang Ye and Pong C Yuen. Purifynet: A robust person re-identification model with noisy labels. *IEEE Transactions on Information Forensics and Security*, 15:2655–2666, 2020.

[42] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 552–561, 2019.

[43] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195, 2020.

[44] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

[45] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9294–9303, 2020.

[46] Anlin Zheng, Yuang Zhang, Xiangyu Zhang, Xiaojuan Qi, and Jian Sun. Progressive end-to-end object detection in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 857–866, 2022.

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[48] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017.

[49] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017.

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

[51] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003.

[52] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.