Proceedings of 2023 IEEE
International Conference on Mechatronics and Automation
August 6 - 9, Harbin, Heilongjiang, China

Pseudo-label Assisted Optimization of Multi-branch Network for Cross-domain Person Re-identification

Zhengyang Wang[1,2], Shuxiang Guo[1,3], Xue Shang[1], and Xiufen Ye[1]*
[1]College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, China
Email: jiollos13@outlook.com, yexiufen@hrbeu.edu.cn, candyxshang@outlook.com
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3]School of Life science, Beijing Institute of Technology, Beijing, China
Email: guoshuxiang@hotmail.com

***Abstract*—Person re-identification plays an important role in the field of robot intelligence perception and public safety. Its main task is to identify person targets under cross-camera. However, the domain diversity between different datasets poses a clear challenge for adapting a model trained on one dataset to another. Currently, person re-identification methods based on domain adaptive learning and pseudo-label have made good progress on this problem. Unfortunately, inferior pseudo-labels and source domain noise affect the performance. In order to improve the quality of generated pseudo-labels and enhance the feature representation capability of the model, we propose a pseudo-label assisted optimization of a multi-branch person re-identification method. The multi-branch network is able to extract and represent more effective global and local features, and the generated pseudo-labels are optimized by using cosine similarity and DBSCAN clustering on the feature vectors, thus improving the consistency of the supervised information to enhance the cross-domain recognition performance. We also use a loss function combining cross-entropy loss and triplet loss to make the best feature learning. Experiments show that our method performs well in the Market-to-Duke and Duke-to-Market cross-domain recognition tasks.**

***Index Terms*—Person re-identification, Pseudo-label, Multi-branch network, Cross domain adaptive**

I. Introduction

Person re-identification (Re-id) is an important research area in the field of pattern recognition and automation, whose main objective is to achieve accurate re-identification of persons [1]. In robotic systems, Re-id also plays an important role. With person re-identification, the robot can realize the recognition, location, and tracking of persons, and then realize the functions of intelligent navigation and human-robot interaction. In addition, re-id can be applied to robot security and monitoring systems to improve the safety and reliability of robots.

Person re-identification aims to accomplish the task of person targets retrieved across cameras. At its core is the problem of instance-level image retrieval across domains [2]. In real situations, the data of the training model and the data to be detected are often not captured by the same camera. For person images from new camera systems, Re-id models trained on existing datasets usually suffer from significant performance degradation due to the domain gap. To address this problem, researchers have used various schemes such as cross-domain adversarial training, transfer learning, data augmentation, and domain adaptive learning. The cross-domain adversarial training based method is usually a Generative Adversarial Networks (GANs) based method [3] [4], which requires a large amount of target domain data for training and requires considerable training time and computational resources, and the model is vulnerable to noise and poor quality generated samples. The transfer learning based method is sensitive to the differences between source and target domain data which requires a deeper understanding of the characteristics of the dataset, otherwise the results will be poor [5] [6]. In contrast, the domain adaptive learning based method enables the model to adapt to new domains by generating pseudo-labels for joint training and iterative optimization. In addition, the method does not require strict annotation of the target domain data, which is relatively easy to implement [7] [8].

However, the domain adaptive learning based method requires the adequate and reasonable selection and processing of samples in the source and target domains, otherwise, the source domain and pseudo-label noise will cause strong interference, resulting in insufficient model generalization capability. Effective feature extraction and high-quality pseudo-labels can greatly reduce the impact of these noises. Among the traditional supervised learning methods, [9] successfully uses multi-branch networks to effectively improve the feature extraction ability. Multi-branch networks can provide better feature representation capability and better feature fusion capability, which can abstract different feature dimensions and improve the generalization performance of the model in unknown domains. Therefore, we propose a multi-branch network based on resnet50 [10] for better representing of global and local features for the characteristics of the person dataset. Besides, among the domain adaptive learning methods, the pseudo-label based domain adaptive learning methods are widely used, but there may be some errors in generating pseudo-labels, so the quality of the pseudo-labels directly determines the performance of model in cross-domain recognition. In this regard, we adopt a pseudo-label optimization method for domain adaptive learning. First, the

This work was supported by the National Natural Science Foundation of China (Grant No. 42276187), the Fundamental Research Funds for the Central Universities, China (Grant No. 3072022FSC0401), and China Scholarship Council (Grant No. 202006680060).

979-8-3503-2084-8/23/$31.00 ©2023 IEEE

feature vector is filtered using cosine similarity during the training process to distinguish irrelevant features and select features that may come from the same person target. Then the feature optimization method based on DBSCAN clustering [11] is used to cluster global and local features in continuous iterations to ensure that all features belonging to the same person in all branches have the same identity and improve the consistency of the supervisory information, which can reduce the error caused by the generated pseudo-labels. At the same time, we also reconstruct the training sample space and design a set of data augmentation strategies to increase the diversity and complexity of the training samples in order to enhance the cross-domain recognition ability of the model.

We experiment and evaluate our method on Market1501 [2] and DukeMTMC-reid [12] dataset, the two most widely used and popular datasets in the field of person re-identification. The experimental results show that our method possesses the ability to effectively improve the performance on cross-domain recognition in person re-identification.

II. Methodology

A. Sample Space Construction

We establish a data mapping to construct a new sample space through a joint data augmentation strategies. Our purpose of this part is to increase the diversity and variability of the training data, which can lead to better generalization and improved performance of the model.

Let Γ be the set of data augmentation operations applied to the original data space. $\gamma_j(x_i)$ represents the image x_i after applying the operation γ_j, including Augmix, Gaussian blur, color jittering, random perspectives, random cropping, and random erasing (mosaic and regular [13]). As shown in Fig. 1

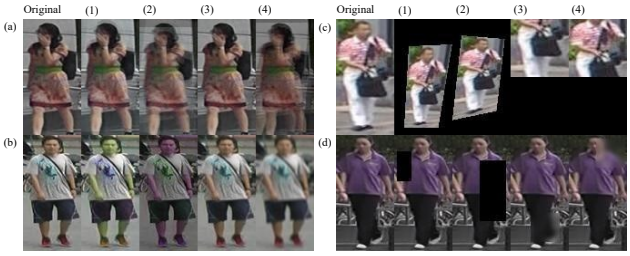

Fig. 1. The Sample space construction for training including Augmix, Gaussian blur, color jittering, random perspectives, random cropping, and random erasing (mosaic and regular)

The transformation can be represented as follows:

$$X^{'} = \{(\gamma_j(x_i), y_i) \mid x_i \in X, \gamma_j \in \Gamma\} \tag{1}$$

where y_i is the label related to x_i. We set a series of parameters $\{t_1, ..., t_k\}$ applying the corresponding transformations γ_j. The new constructed sample space is also defined at the identity-level, which means the transformed image is corresponding to a true identity of the original one. By constructing new training sample spaces that are similar but not identical to the original data spaces, the model can learn to recognize persons under a wide range of conditions and generalize better to unseen data.

B. Network Structure

We propose a multi-branch network for this task. The network architecture proposed applies the ResNet50 backbone to obtain a feature map f. f represents the learned features extracted from the image at different levels of abstraction. After the res_conv4_1 block, this feature map is split into three branches: a global branch that applies global average pooling to obtain a global feature vector f_{global}, and two local branches that divide the feature map into horizontal parts and apply average pooling to obtain several fine-grained feature vectors. The feature space has dimensions equal to the total number of features extracted from the Resnet50 backbone and the three branches, which is the sum of the dimensions of the global feature vector f_{global} and the local feature vectors f_{local1} and f_{local2}. Resnet50 is set as a backbone to extract features. After the res_conv4_1 block, the backbone feature maps are split into three different branches:

Global Branch:

$$f_{\text{global}} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} f_{i,j} \tag{2}$$

where H and W are the height and width of the feature map f. $f_{i,j}$ corresponds to the response of the feature map f at the spatial location (i, j) to a particular feature or pattern in the input image. By computing the average activation across all spatial locations of f, we obtain the global feature vector f_{global}, which summarizes the most salient features of the input image, while the local branches extract more fine-grained features from different regions of the feature map.

Local Branch 1:

We split the feature maps obtained from the backbone into two parts along the horizontal direction, concatenate them, and apply a global pooling layer on the concatenated feature map:

$$\begin{aligned} f_{\text{local1}} = &\frac{1}{2HW} \sum_{i=1}^{H/2} \sum_{j=1}^{W} f_{i,j} \\ &\oplus \frac{1}{2HW} \sum_{i=H/2+1}^{H} \sum_{j=1}^{W} f_{i,j} \end{aligned} \tag{3}$$

Local Branch 2:

Similar to Local Branch 1, we split the feature maps obtained from the backbone into six parts along the horizontal direction, concatenate them, and apply a global pooling layer on the concatenated feature map:

$$\begin{aligned} f_{\text{local2}} = &\frac{1}{6HW} \sum_{i=1}^{H/6} \sum_{j=1}^{W} f_{i,j} \\ &\oplus ... \oplus W_{l2} \frac{1}{6HW} \sum_{i=5H/6+1}^{H} \sum_{j=1}^{W} f_{i,j} \end{aligned} \tag{4}$$

Finally, we concatenate the global feature vector and the feature vectors from Local Branch 1 and Local Branch 2 to get the final feature vector:

$$f_{\mathrm{o}} = F(x) = concat(f_{global}, f_{local1}, f_{local2}) \tag{5}$$

where $F(x)$ refers to the function that maps an input image x to a output feature map f_o, which encodes the salient features of the image across different levels of abstraction. The structure of our proposed network structure is shown in Fig. 2.

C. Pseudo-label Generation

For pseudo-label-based domain adaptive person re-identification, the implementations are usually divided into two parts. The first part is to generate the pseudo-label. The second part is to improve the quality of the generated pseudo-label, in our paper, specifically is to optimize the quality and the consistency of all the supervised information during training. This first part is carried out in the following way:

- Train a pre-trained model using source data.
- Predict these samples from the target domain.
- Generate pseudo-label for the selected samples from target domain.

Firstly, we define the source data as $D_s = \{(x_i, y_i)\}_{i=1}^{N_s}$, in which x_i denotes the i-th sample image, and y_i denotes the i-th label related to the x_i. Similarly, $D_t = (x_j)_{j=1}^{N_t}$ denotes the target data. The proposed resnet-based multi-branch network is used for feature extraction as shwon in (5). In the pre-training part, the extracted feature can be presented as $f_s(x_i)$, where $i \in [1, N_s]$.

A cross-entropy loss is used to measure the difference between the predicted label and the true label:

$$\mathcal{L}_{\mathrm{ce}} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{k=1}^{K} y_{i,k} \log\left(p_{i,c}\right) \tag{6}$$

where K is the number of categories, $y_{i,c}$ denotes the true label of the kth category of the ith image, and $p_{i,k}$ denotes the predicted probability of the kth category of the ith image.

To reduce the distribution differences between the source and target domains, pseudo-label is used to assist the training. Specifically, a pre-trained model is used to extract features from the target domain data to obtain the feature vector $f_t(x_i)$, where $i \in [1, N_t]$. Then, the softmax function is used for $f_t(x_i)$ to obtain the prediction probability vector $p_t(x_i)$:

$$p_{\mathrm{t,\,i,\,k}} = \frac{\exp\left(w_k^T f_t\left(x_i\right)\right)}{\sum_{k=1}^{K} \exp\left(w_k^T f_t\left(x_i\right)\right)} \tag{7}$$

where w_k is the weight vector of the softmax function, $k \in [1, K]$. The $p_t(x_i)$ is used as the pseudo label of the target domain data, and the target domain data is added to the training using the pseudo label method. Specifically, the target domain data and pseudo label are formed into a new dataset $D' = (x_i, p_t(x_i))_{i=N_s+1}^{N_s+N_t}$, and then D' is trained together with the source domain dataset D_s. The cross-entropy loss function is used to measure the difference between the predicted labels and the pseudo-labels:

$$\mathcal{L}_{\mathrm{ce\_pre}} = -\frac{1}{N_s + N_t} \sum_{i=1}^{N_s+N_t} \sum_{k=1}^{K} y_{i,k} \log\left(p_{i,k}\right) \tag{8}$$

where $y_{i,k}$ denotes the true label or predicted label of the kth category of the ith image, and $p_{i,k}$ denotes the predicted probability of the kth category of the ith image.

In addition, to further improve the robustness and generalization performance of the model, triplet loss can be used to train the model. Specifically, for each image x_i in the source and target domains, its feature vector $f(x_i)$ can be used to calculate the distance d_{ij} between it and other images $f(x_j)$ of the same identity and the distance d_{ik} between it and other images $f(x_k)$ of different identities, and then triplet loss can be used to encourage images to be closer together and images with different identities to be further apart, which is:

$$\begin{aligned} \mathcal{L}_{\mathrm{tri\_pre}} = \frac{1}{N_t} \sum_{i=N_s+1}^{N_s+N_t} \sum_{a,p,n} \max\left(0, d_{ia} - d_{ip} + m\right) \\ + \max\left(0, d_{ia} - d_{in} + m\right) \end{aligned} \tag{9}$$

where a, p, n denotes three images of the same identity for x_i and m is the margin parameter that controls the gap between the distance between images of the same identity and the distance between images of different identities.

For the pseudo-label generation step, which is closely related to the second part, the generated pseudo-label is also updated when the quality of the pseudo-label and the consistency of the supervised information are optimized.

D. Optimization for the quality of pseudo-label and the consistency of supervised information

For cross-camera person re-identification tasks, there is a domain gap between different datasets. The quality of pseudo-label is the most important part. The way in our paper to improve it is to make all the supervised information in the training process be highly consistent, so that the learned model is of higher robustness.

Firstly, we calculate the cosine similarity score between each sample from source domain $D_s = \{(x_i, y_i)\}_{i=1}^{N_s}$ and target domain $D_t = (x_j)_{j=1}^{N_t}$. The cosine similarity measures the degree of similarity between two feature vectors:

$$d(f_s(x_i), f_t(x_j)) = \frac{f_s(x_i) \cdot f_t(x_j)}{\|f_s(x_i)\| \, \|f_t(x_j)\|} \tag{10}$$

where $d(f_s(x_i), f_t(x_j))$ denotes the distance of two feature vectors, $\cdot$ denotes the vector dot product and $\|\cdot\|$ denotes the L_2 parametrization of the vector.

The similarity threshold τ_1 is set fixed throughout the training process. The results can be represented as follows:

$$\mathcal{S} = \{f_t(x_j) \mid d(f_s(x_i), f_t(x_j)) \geq \tau_1, j \in [1, N_s]\} \tag{11}$$

where $\mathcal{S}$ denotes the set of feature vectors from source domain with similarity scores above a threshold τ_1 to the feature vectors $f_t(x_j)$ from target domain, the pseudo-label set related

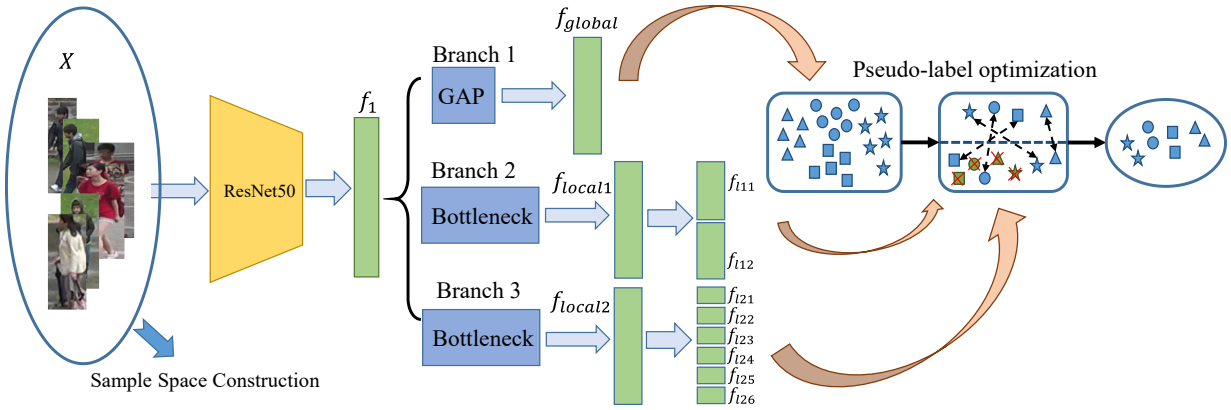

Fig. 2. The structure of our proposed network. There are three branches for feature extraction. After the backbone, there is a pseudo-label optimization stage.

to $\mathcal{S}$ can be defined as $Y_{\text{pseudo}} = \{y_1, y_2, ..., y_m\}$ for the first stage, where $y_i \in \mathbb{R}^d$ denotes the pseudo-label of the ith sample.

Then, Y_{pseudo} are the input of clustering optimization, which is the second stage. Here we use the Density-based spatial clustering of applications with noise (DBSCAN) to cluster the samples. DBSCAN is a density clustering algorithm whose core idea is to divide high-density regions into a cluster and low-density regions into noise. The clustering process is performed independently in each iteration. The output is a set of clusters $\mathcal{C} = \{C_1, C_2, ..., C_k\}$.

In the clustering processing, we try to optimize the Y_{pseudo} into a set of clusters $\mathcal{C}$. Specifically, let the distance matrix be $D \in \mathbb{R}^{m \times m}$, where m is the number of pseudo-labels. For each sample $y_i \in Y_{\text{pseudo}}$, compute its domain N_i in D and count the number of samples it contains $|N_i|$, as (12):

$$N_i = \{y_i \in Y_{\text{pseudo}} \mid D_{ij} \leq \epsilon\}, \quad |N_i| \geq \text{minPts} \tag{12}$$

where ϵ is the distance threshold parameter in DBSCAN and minPts is the minimum number of fields for the core objects in DBSCAN.

For all samples y_i marked as core objects, add them and all their samples in N_i to a cluster $C_p \in \mathcal{C}$. For all samples y_j that are marked as non-core objects, mark them as boundary points and remove them from Y_{pseudo}. For each cluster C_p, if the number of samples it contains is less than a threshold minSize, it will be deleted. The final remaining clusters $\mathcal{C} = \{C_1, C_2, ..., C_k\}$ are the optimized pseudo-label clustering results. The algorithm of our optimization method is shown in Algorithm 1.

E. Loss Function

The total loss function in the optimization process can be represented as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{tri} \tag{13}$$

where λ_1 and λ_2 are the loss weight for two kinds of loss functions. $\mathcal{L}_{ce}$ denotes the cross-entropy loss in the source and target domains for training the classifier, $\mathcal{L}_{tri}$ denotes the triplet loss for training the feature extractor, and $\mathcal{L}_p$ denotes the pseudo-label loss for training the classifier and optimizing the performance of domain adaptation. The $\mathcal{L}_{ce}$ and $\mathcal{L}_{tri}$ are set the same as (8) and (9).

III. Experiments and Results

A. Dataset and Metrics

We evaluate our proposed method on two widely used person re-identification datasets: Market1501 and DukeMTMC-reid. The Maket1501 dataset has 32,668 images containing 1501 identities, of which 12,936 images in the training set contain 751 identities, obtained by 6 camera. The DukeMTMC-reid dataset has 36,411 rows of images containing 702 identities, of which 16,552 images in the training set, captured by 8 cameras. We conduct domain adaptive experiments on these three datasets, only data from the source domain provided label information. Mean average precision (mAP) and Cumulative Catch Characteristics Curve (CMC) top-1 accuracy are adopted to evaluate the model performance.

B. Implementation Details

We use a multi-branch resnet50 based network as the backbone. We design a set of data augmentation strategies for sample space construction, which makes the training samples more diverse and thus can improve the generalization of model performance. The data augmentation strategies are used in both the pre-training and optimization processes, which include Augmix, Gaussian blur, color jittering, random perspectives, random cropping, and random erasing (mosaic and regular). All images are resized to 256×128 before being sent into the backbone network. We randomly sample 4 instances per ground truth (in pre-training) and pseudo label (in optimization) in a mini-batch, resulting in batch-size 64. Adam optimizer is used in all the training process with a weight decay of 0.0005. For the pre-training stage, the initial learning rate is set to 0.00035 and is decreased to 0.1 of its previous value on the 40-th and 70-th epoch in the total 100 epochs. For the optimization stage, the learning rate is fixed to 0.00035 for overall 40 training epochs. We adopted loss

Algorithm 1: Optimization for Pseudo-label and Consistency of Supervised Information

Input: Source domain data $D_s = \{(x_i, y_i)\}_{i=1}^{N_s}$, target domain data $D_t = (x_j)_{j=1}^{N_t}$
Output: Optimized pseudo-label clustering results

Step 1: Calculate cosine similarity scores;
for $i = 1$ **to** N_s **do**
  for $j = 1$ **to** N_t **do**
    Calculate $d(f_s(x_i), f_t(x_j))$ as defined in Eq. (10);
    if $d(f_s(x_i), f_t(x_j)) \geq \tau_1$ **then**
      Add $f_t(x_j)$ to $\mathcal{S}$ as defined in Eq. (11);
    end
  end
end
Generate pseudo-label set Y_{pseudo} from $\mathcal{S}$;
Step 2: Cluster optimization using DBSCAN;
Initialize D with pairwise distances between samples in Y_{pseudo};
for $i = 1$ **to** m **do**
  Compute N_i and $|N_i|$ as defined in Eq. (12);
  Mark y_i as a core object if $|N_i| \geq$ minPts, otherwise mark as a non-core object;
  if y_i *is a core object* **then**
    Add y_i and all samples in N_i to a new cluster C_p;
  end
  else
    Mark y_i as a boundary point and remove it from Y_{pseudo};
  end
  $\mathcal{C} = \bigcup \mathcal{C}_p$
end
Remove clusters with less than minSize samples;
return $\mathcal{C} = \{C_1, C_2, ..., C_k\}$;

function in (13), in which λ_1=0.5, λ_2=0.8. The m of triplet loss is set to m=0.3. For the optimization stage, ϵ is set to 0.5 and minPts is set to 4. We set two tasks to evaluate our method: Market1501-to-DukeMTMC-reid and DukeMTMC-reid-to-Market1501. The baseline is a single branch resnet50 based network.

C. Results and Analysis

The results of our experiments for domain adaptive in task Market1501-to-DukeMTMC-reid (Market-to-Duke) and DukeMTMC-reid-to-Market1501 (Duke-to-Market) are shown in Table I. Three typical unsupervised learning method for person re-identification methods are also listed for comparison.

As we can see from Table I, our method is also superior to the highest of the three typical methods in mAP by 18.4% (Market-to-Duke), 16.9% (Duke-to-Market), and also in CMC (top-1) by 10.1% (Market-to-Duke) and 10.5% (Duke-to-Market).

TABLE I
RESULTS AND COMPARISON

Methods	Market-to-Duke		Duke-to-Market	
	mAP	top-1	mAP	top-1
ENC [14]	40.4	63.3	43.0	75.1
PCB-PAST [15]	54.3	72.4	54.6	78.4
SSG [7]	53.4	73.0	58.3	80.0
Pre-trained only	21.3	46.2	28.5	59.7
Ours	**72.7**	**83.1**	**77.2**	**93.5**

we evaluate each component of the proposed method. Compared baselines are as follows:

- **Pre-trained only**: Pre-trained only without optimization and fine-tuning stage.
- **Final model (w/o DG)**: Final model only without the data augmentation (AA) part.
- **Final model (w/o L_{tri})**: Final model only without a triplet loss in the whole training stage.
- **Final model (w/o L_{ce})**: Final model only without a cross-entropy loss in the whole training stage.
- **Original ResNet50**: The backbone is a single branch original resnet50 network, only optimizing and fine-tuning on global features.

The ablation studies are shown in Table II and Table III. When our multi-branch network is adopted, the performance is significantly better than with the original resnet50: 8.0% (Market-to-Duke) and 8.4% (Duke-to-Market) better on mAP, and 4.8% and 4.6% better on CMC (top-1).

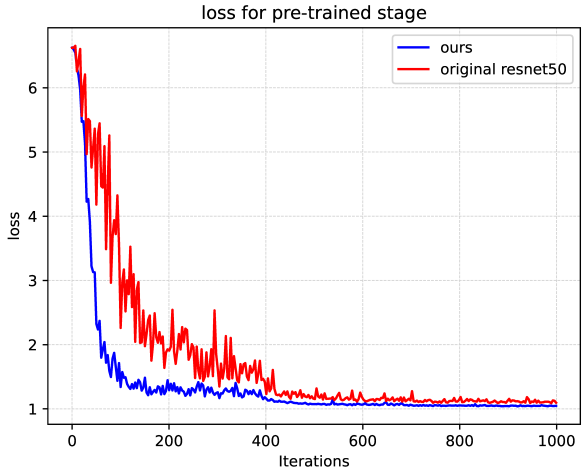

Fig. 3. The loss curse for pre-training stage of our method and baseline model. The red line denotes the baseline, and the blue line denotes ours.

The loss curse for pre-training stage is shown in Fig. 3. From the Fig. 3, we can obviously observe that the difference in performance between the two networks can be illustrated by the magnitude of the drop and the jitter of the two curves. Our network drops faster and has less jitter during the drop. Our network has stronger fitting ability and better performance for feature extraction and recognition.

TABLE II
ABLATION STUDIES ON MARKET-TO-DUKE TASK

Market-to-Duke	Original ResNet50		Our Backbone	
	mAP	top-1	mAP	top-1
Pre-trained only	18.4	32.5	21.3	46.2
Final model (w/o AA)	57.9	73.8	68.2	86.5
Final model (w/o L_{tr})	59.6	76.3	68.6	86.2
Final model (w/o L_{ce})	59.1	75.5	66.9	81.0
Final model	**64.7**	**78.3**	**72.7**	**83.1**

TABLE III
ABLATION STUDIES ON DUKE-TO-MARKET TASK

Duke -to-Market	Original ResNet50		Our Backbone	
	mAP	top-1	mAP	top-1
Pre-trained only	20.1	39.7	28.5	59.7
Final model (w/o DG)	65.5	79.6	69.3	87.6
Final model (w/o L_{tr})	63.7	77.1	72.2	89.2
Final model (w/o L_{ce})	62.1	77.4	71.5	88.1
Ours	**67.8**	**85.9**	**77.2**	**93.5**

Compared with the 'pre-trained only', our model with the optimization stage has stronger domain adaptive recognition capability and performs better in both Market-to-Duke and Duke-to-Market tasks.

In addition, We use a loss function with a mixture of cross-entropy and triplet loss. During the training process, cross-entropy loss and triplet loss can promote each other, thus improving the performance of the model. Specifically, Cross-entropy loss is used to improve the classification accuracy of the model, and triplet loss is used to enhance the model's ability to distinguish between persons. Cross-entropy loss can make the model converge quickly and achieve high classification accuracy in the initial stage, while triplet loss can enhance the discrimination between samples in the later stage of training, thus improving the robustness and generalization ability of the model. It can also be seen from Tables II and Table III that using only one loss will not achieve the best performance, and the performance of the model will both be degraded.

IV. CONCLUSION

In this paper, we propose a multi-branch person re-identification method with pseudo-label assisted optimization method. Our method achieves good performance in cross-domain recognition. We propose a multi-branch network that can better extract and represent global and local features. The input training data is also reconstructed in sample space to make it more diverse and complex, which can enhance the generalization of the model. We also propose a pseudo-label optimization method to improve the quality of the generated pseudo-labels, which consists of cosine similarity and DBSCAN clustering. This method can effectively improve the pseudo-label quality and the consistency of the supervised information. In the training process we use a combination of cross-entropy loss and triplets loss to maximize the cross-domain learning ability of the model. Ultimately, our method shows good performance in both Market-to-Duke and Duke-to-Market cross-domain tasks.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 42276187), the Fundamental Research Funds for the Central Universities, China (Grant No. 3072022FSC0401), and China Scholarship Council (Grant No. 202006680060).

REFERENCES

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[2] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015.

[3] S. Zhou, M. Ke, and P. Luo, "Multi-camera transfer gan for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 393–400, 2019.

[4] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 79–88, 2018.

[5] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1306–1315, 2016.

[6] Y.-J. Li, F.-E. Yang, Y.-C. Liu, Y.-Y. Yeh, X. Du, and Y.-C. Frank Wang, "Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 172–178, 2018.

[7] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *proceedings of the IEEE/CVF international conference on computer vision*, pp. 6112–6121, 2019.

[8] Y. Zou, X. Yang, Z. Yu, B. V. Kumar, and J. Kautz, "Joint disentangling and adaptation for cross-domain person re-identification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 87–104, Springer, 2020.

[9] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang, "Multiple granularity descriptors for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision*, pp. 2399–2406, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[11] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "Dbscan: Past, present and future," in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pp. 232–238, IEEE, 2014.

[12] M. Gou, S. Karanam, W. Liu, O. Camps, and R. J. Radke, "Dukemtmc4reid: A large-scale multi-camera person re-identification dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 10–19, 2017.

[13] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 13001–13008, 2020.

[14] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 598–607, 2019.

[15] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8222–8231, 2019.