CUPR: Contrastive Unsupervised Learning for Person Re-identification

Khadija Khaldi[a] and Shishir K. Shah[b]
Quantitative Imaging Laboratory, Department of Computer Science, University of Houston, U.S.A.

Keywords: Person Re-identification, Unsupervised Learning, Contrastive Learning, Deep Learning.

Abstract: Most of the current person re-identification (Re-ID) algorithms require a large labeled training dataset to obtain better results. For example, domain adaptation-based approaches rely heavily on limited real-world data to alleviate the problem of domain shift. However, such assumptions are impractical and rarely hold, since the data is not freely accessible and require expensive annotation. To address this problem, we propose a novel pure unsupervised learning approach using contrastive learning (CUPR). Our framework is a simple iterative approach that learns strong high-level features from raw pixels using contrastive learning and then performs clustering to generate pseudo-labels. We demonstrate that CUPR outperforms the unsupervised and semi-supervised state-of-the-art methods on Market-1501 and DukeMTMC-reID datasets.

1 INTRODUCTION

Person re-identification (re-ID) is an important task in computer vision that aims to match a person across camera views. Significant research from both academia and industry has been done to address this problem. Over the past decade, most of the existing methods mainly focus on hand-crafted algorithms (Gray and Tao, 2008; Farenzena et al., 2010), salience analysis (Zhao et al., 2013; Wang et al., 2014), and dictionary learning (Liu et al., 2014). With recent advances in deep learning and the rising demand for intelligent video surveillance, this problem has also been addressed using convolutional neural network (CNN) models. These methods are categorized into supervised, domain adaptation, and unsupervised learning. While, extensive work has been done in supervised representation learning (Fu et al., 2019; Geng et al., 2016; Sun et al., 2018; Li et al., 2018b) to improve the performance of person re-ID, they fundamentally require massive labeled data which is unfeasible and expensive to acquire. Therefore, domain adaptation and unsupervised learning approaches were proposed upon the success of deep learning (Wei et al., 2018; Deng et al., 2018; Wang et al., 2018; Lin et al., 2019). An overview of the previous architectures is shown in Figure 1. While domain adaptation methods aim typically to learn a discriminative representation by pre-training the model on labeled source data and then adapt it to the unlabeled target data. With limited knowledge of the overlap of the source and the target distribution and the significant gap between them, these methods remain impractical. To address the previous problems, end-to-end unsupervised methods propose to take advantage of unlabeled data, e.g. bottom-up clustering approach (Lin et al., 2019). Specifically, the latter model starts by considering each sample as a single cluster, then merges the clusters by applying a bottom-up clustering approach on the extracted features. While it does not require annotated data, this method highly depends on the clustering quality, which can result in degrading the performance due to incorrect pseudo labels. To tackle this challenge, we propose CUPR – Contrastive Unsupervised Learning for Person re-identification, a novel pure unsupervised learning method. CUPR uses a form of contrastive learning that maximizes agreement between positive pair of the same observation, generated by a composition of data augmentation, as shown in Figure 2. We show that CUPR significantly outperforms prior unsupervised Re-ID state-of-the-art methods by performing contrastive learning simultaneously with clustering in an iterative fashion. This allows us to learn strong feature embedding before generating pseudo labels in a very simple way requiring no specialized architectures.

Our paper makes the following key contributions: We present CUPR, a simple pure unsupervised framework that integrates contrastive learning with clustering. We empirically show that contrastive learn-

[a] https://orcid.org/0000-0001-8845-1444
[b] https://orcid.org/0000-0003-4093-6906

Khaldi, K. and Shah, S.
CUPR: Contrastive Unsupervised Learning for Person Re-identification.
DOI: 10.5220/0010239900920100
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 92-100
ISBN: 978-989-758-488-6

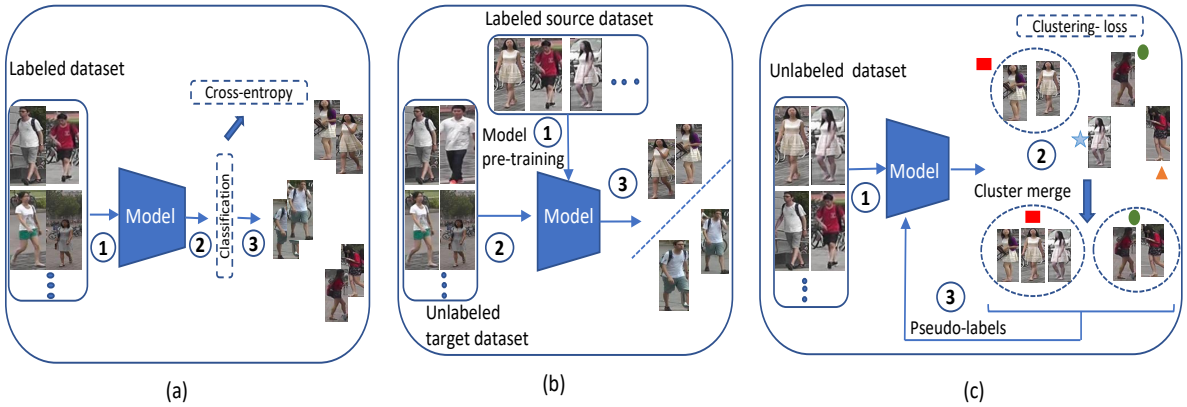

Figure 1: Demonstration of previous person Re-ID methods. (a) Supervised methods, (b) Unsupervised cross-domain adaptation and (c) Bottom-up image clustering.

ing can improve the performance of person re-ID, without using any complicated pipelines as the previous methods suggest. The experimental results demonstrate that our approach outperforms the state-of-art methods on two large-scale datasets including Market-1501 (Zheng et al., 2015) and DukeMTMC-reID (Gou et al., 2017).

2 RELATED WORK

Most existing person re-id methods are based on supervised learning, they mainly focus on metric learning (Geng et al., 2016; Hermans et al., 2017; Lisanti et al., 2014; Zhong et al., 2017) and view-invariant feature learning (Deng et al., 2018; Li et al., 2018b; Bian et al., 2019). However, their scalability is very limited in the real world, where collecting a tremendous amount of labeled data is expensive and infeasible. To alleviate the above limitation, unsupervised person Re-ID methods are proposed to learn from unlabeled data without expensive manual annotation. Most of them can be considered as a deeply unsupervised approach, an end-to-end unsupervised approach, or an unsupervised domain adaptation approach.

2.1 Deeply Unsupervised Person Re-ID

For deeply unsupervised methods, cross-camera label estimation is one of the approaches. Ye *et al.* propose a dynamic graph matching method (Ye et al., 2017) where the estimated labels and learned metric are updated in an iterative manner. Liu *et al.* perform K-reciprocal nearest neighbor search and negative mining (Liu et al., 2017) to estimate the identities of training tracklets. Similarly, Ye *et al.* iteratively assign labels to the unlabeled sequences via robust anchor embedding and top-k counts label prediction, to collect more anchor video sequences (Ye et al., 2018).

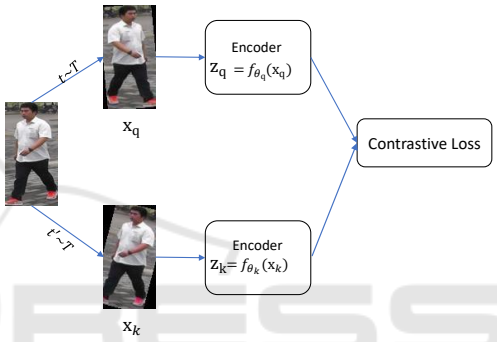

Figure 2: Demonstration of contrastive learning for person re-identification. Two random data augmentation (t∼T and t'∼ T) are applied to an anchor to generate positive pair. The input sample is treated as an anchor while x_k and x_q is the positive pair. CUPR trains an encoder network $f_\theta(.)$ to ensure that data-augmented embeddings of the same identity are close to each other.

2.2 End-to-End Unsupervised Person Re-ID

Li *et al.* propose a Tracklet Association Unsupervised Deep Learning (TAUDL) framework (Li et al., 2018a), which jointly conducts the within-camera tracklet discrimination and models the cross-camera tracklet association. Similarly, Wu *et al.* introduce an unsupervised camera-aware similarity consistency mining approach (Wu et al., 2019) by exploring the relation of pairwise similarity between intra-camera matching and cross-camera matching. Yang *et al.* propose a PatchNet model (Yang et al., 2019) to learn discriminative patch features. They train their model by pulling the features of similar patches together and pushing dissimilar ones away.

On the other hand, some methods explore iterative clustering to reduce the labeling efforts. Lin *et al.* iteratively train their model using pseudo-labels generated by bottom-up clustering (Lin et al., 2019).

Our method aligns with the last approach, as we are using an iterative clustering approach. However, we propose to learn a strong adaptative feature embedding using contrastive loss, before generating pseudo-labels.

2.3 Unsupervised Domain Adaptation

Recently, cross-domain transfer learning is used in the unsupervised re-ID task where knowledge from an external labeled source dataset is transferred to an unlabeled target dataset.

Among these existing works, GAN has been used to transfer the source domain images to target-domain style. To handle the domain shift problem, Wei *et al.* (Wei et al., 2018) propose a Person Transfer Generative Adversarial Network (PTGAN), transferring the knowledge from one labeled source dataset to another unlabeled target dataset. Similarly, a "learning via translation" framework (Deng et al., 2018) is proposed with similarity preserving image generation (SPGAN). Also, Wang *et al.* develop TJ-AIDL framework (Wang et al., 2018), which learns jointly an attribute semantic and identity discriminative feature representation space from a labeled source domain, then transfer it to any an unlabeled domain.

Similar to supervised learning, these domain adaptation approaches suffer from the need for collecting annotations for the source dataset. Thus, our work focuses on the pure unsupervised setting, where there is no need for data annotation.

3 PROPOSED METHOD

In this section, we introduce our learning framework CUPR for unsupervised person re-identification. We first present the contrastive learning approach, a key component of our model, which is able to learn strong representations with no labels. Then, we perform a hierarchical-based clustering algorithm on the learned features to generate pseudo-labels. In principle, one could use any clustering algorithm in the CUPR pipeline. The overview of our proposed framework is described in Figure 3.

Given a training set:

$$D = \{x_1, x_2, \ldots, x_N\} \tag{1}$$

of N unlabeled images, our goal is to learn an encoder $f_\theta(.)$ that maps high dimensional pixels to strong feature vector V_i.

$$V_i = f_\theta(x_i) \tag{2}$$

CUPR is mainly composed of a CNN backbone and a pseudo-label generation network. To optimize the model without human supervision, we propose two types of self-supervision: (1) contrastive learning ($\mathcal{L}_c$) and (2) clustering-based learning ($\mathcal{L}_{cl}$). Combining these two losses, the model is able to jointly consider generating rich representation and forming well-separated and dense clusters.

At evaluation time, the trained encoder $f_\theta(.)$ generate feature vectors for the query set $\{x_1^q, x_2^q, \ldots, x_{N_q}^q\}$ and the galley set $\{x_1^g, x_2^g, \ldots, x_{N_g}^g\}$. Then, a pairwise distance defined as following, is used to perform re-id matching.

$$D(x_i^q, x_i^g) = \|f_\theta(x_i^q) - f_\theta(x_i^g)\| \tag{3}$$

3.1 Contrastive Stage

In recent years, discriminative approaches have shown a lot of promise in learning visual representation without human supervision, particularly accelerated by a well-known paradigm called contrastive learning (Hadsell et al., 2006; Dosovitskiy et al., 2014; Oord et al., 2018; Bachman et al., 2019).

One of the key challenges for contrastive learning is defining positives and negative samples relative to an anchor. Inspired by recent contrastive learning algorithms (Chen et al., 2020; He et al., 2019), for each data sample we create two positive pair (x_i and x_j), by applying a composition of multiple data augmentation operations. This will generate images with a rich appearance and a view variation. In this work, we sequentially apply three simple augmentations: random cropping, horizontal flipping, and color jittering. Other data augmentation combination were used, but they resulted in lower performance. Then, our base encoder $f_\theta(.)$ extracts feature vectors form the augmented data examples as following.

$$f_\theta(x_i) = z_i \quad , z_i \in \mathbb{R} \tag{4}$$

To this end, we randomly sample a mini-batch of M examples and generate a pair of positive pair for each sample, resulting in 2M data points. Differently from SimCLR (Chen et al., 2020), we explicitly define our negative samples. Otherwise, while trying to maximize the agreement between the augmented views, we will minimize the agreement between them and other samples of the same person identity. Therefore, to avoid this problem we use a neighborhood approach to mine the negative samples in a mini-batch which is illustrated in Figure 4. Given the feature vectors z_i, we compute the pairwise distance between z_i

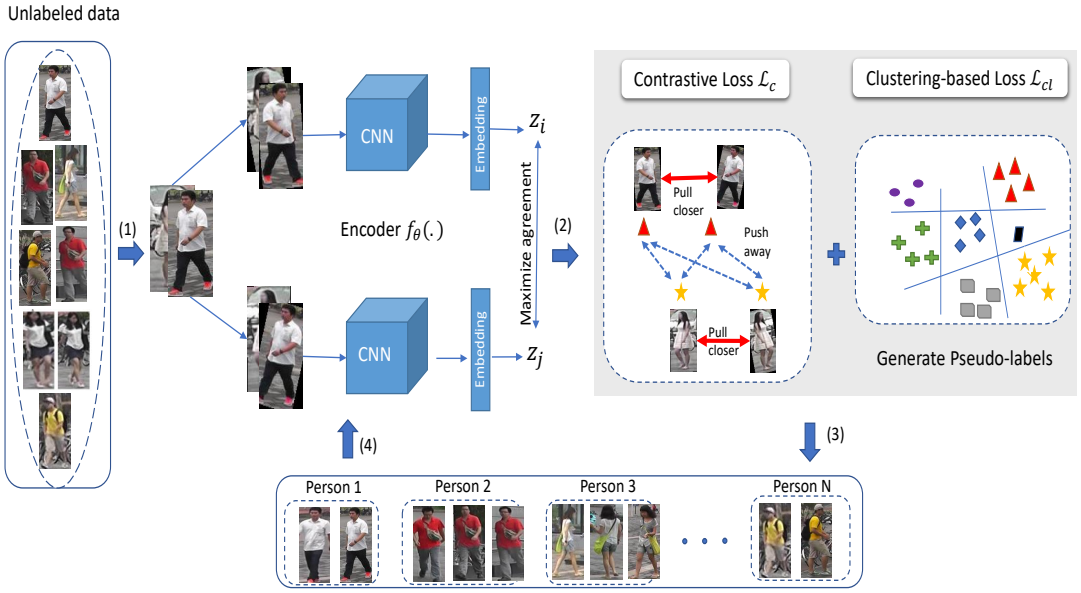

Figure 3: An illustration of the Contrastive Unsupervised Learning for Person Re-Identification. CUPR is an iterative framework mainly composed of a CNN backbone and aims to learn rich representations. First, our framework extracts image features with a CNN model, then the HDBSCAN clustering method is performed using the feature similarities to generate pseudo-labels. In an iterative end-to-end fashion, our model CUPR is trained using two optimization functions: (1) contrastive loss and (2) clustering-based loss.

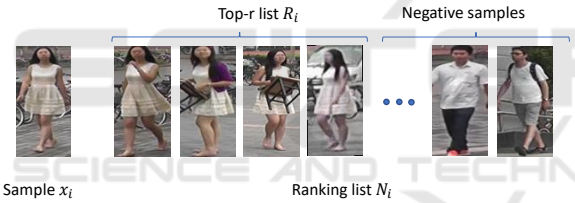

Figure 4: An illustration of mining negative samples in a mini-batch. We compute the ranking list N_i for the target image x_i, then we select negative samples not belonging to top-r list R_i.

and all other augmented examples z_k within a mini-batch, and thus we can get the ranking list N_i for the sample x_i. Then, we argue that the augmented samples not in the top-r nearest neighbor list R_i are likely to be the negative samples.
Finally, we define our contrastive loss function for a positive pair (i, j) in a mini-batch as follows:

$$\mathcal{L}_c = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1, k\notin R_i}^{2M} \mathbb{1}_{[k\neq i]} \exp(sim(z_i, z_k)/\tau)} \quad (5)$$

where $\mathbb{1}_{[k\neq i]}$ is an indicator function evaluating to 1 iff $k \neq i$ and τ is a temperature parameter that controls the softness of probability distribution over classes. The final loss is computed across all positive pairs, both (i, j) and (j, i), in a mini-batch.

3.2 Clustering Stage

Combining clustering and representation learning is one of the most effective strategies for unsupervised learning. This has been adopted by the recent DeepCluster method (Caron et al., 2018). However, it is not trivial to apply clustering for person re-id due to the presence of different viewpoints and low-image resolution. Therefore, we design a clustering-based learning strategy that not only generates reliable clusters but also boosts the model performance.

Given the training data $T = \{x_1, x_2, \ldots, x_N\}$, at the beginning we generate feature vectors $F = \{f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_N)\}$, then we compute the cosine similarity between the feature vectors of sample x_i and all the other samples to generate the following distance matrix D as:

$$D = [D(x_1)\ D(x_2)\ \ldots\ D(x_N)]^T, \quad (6)$$

$$D(x_i) = V^T.v_i, \quad v_i = f_\theta(x_i)$$
$$\forall i \in \{1,\ 2,\ \ldots,\ N\}$$

Then, we apply a hierarchical density-based clustering algorithm (HDBSCAN) (Campello et al., 2013) (other clustering methods can be employed) to obtain reliable clusters for self-supervision. Each cluster is considered as a specific class, in which samples of the same cluster can be assigned to the same pseudo label.

Algorithm 1: Contrastive Unsupervised Learning for re-ID.

Input : Unlabeled training data $T = \{x_i\}^N$; I_{max}; $epoch_{max}$; $batch_{max}$
Output : Model M.
Initialization: Initialize Model parameters θ

```
for step = 1 → I_max do
    for e = 1 → epoch_max do
        for b = 1 → batch_max do
            Generate positive pair z_i and z_j
            Extract feature vectors V = f_θ(.)
            Compute distance matrix Eq.(6)
            Compute contrastive-loss L_c Eq.5
            Compute cluster-loss L_cl Eq.(7)
            Backward to update θ Eq.(9)
            Update the memory bank V_C
        end for
    end for
    Generate new pseudo-labels
    Initialize memory V with new dimensions
    Evaluate the performance of f_θ(.)
    if Perf > Perf* then
        θ* = θ
    end if
end for
```

Note that HDBSCAN discard images not belonging to any cluster by considering them as noise. To address this limitation, we consider discarded images as individual clusters of single data points.

Finally, we minimize the clustering-based loss function $\mathcal{L}_{cl}$, which is formulated as follows:

$$\mathcal{L}_{cl} = -\log \frac{\exp(V_{c,i}^T v_i/\tau)}{\sum_{j=1}^{C} \exp(V_{c,j}^T v_i/\tau)} \quad (7)$$

where V_c is an external memory bank that stores the feature vectors for each cluster, and C is the number of clusters at the current step.
At the first training stage, $C = N$. τ denotes a temperature parameter. The memory bank is updated as:

$$V_{y_i}(t) \leftarrow \frac{1}{2}(V_{y_i} + V_{y_i}(t-1)) \quad (8)$$

where V_{y_i} denotes the up-to-date y_i-th column of the memory bank V.

To summarize, the total loss function for each image in a mini-batch of our model is then formulated as:

$$\mathcal{L} = \mathcal{L}_{cl} + \lambda \mathcal{L}_c \quad (9)$$

4 EXPERIMENT

4.1 Datasets

To evaluate our proposed method, we carried out experiments on two large-scale person Re-ID datasets, namely Market-1501 (Zheng et al., 2015), DukeMTMC-reID (Gou et al., 2017). Market-1501 (Zheng et al., 2015) is a large-scale dataset for person re-ID captured by 6 cameras on a university campus, where pedestrians are detected and cropped by the Deformable Part Model (DPM) (Felzenszwalb et al., 2009). The dataset contains 12,936 images of 751 identities for training and 19,732 images of 750 identities for testing. DukeMTMC-reID (Gou et al., 2017) is a large-scale re-ID dataset derived from the DukeMTMC dataset (Ristani et al., 2016) and has 8 cameras. It contains 36,411 labeled images of 1,404 identities. Similar to Market1501, 702 identities are used for training and remaining identities as testing.

4.2 Protocols

For performance measurement, we use the cumulative matching characteristic (CMC) and the mean Average Precision (mAP). We report the rank-k scores which represent the retrieval precision and the mAP value which reflects the overall recall rate.

4.3 Implementation Details

We adopt ResNet-50 (He et al., 2016) without the last classification layer as the base neural network encoder with pre-trained weights on ImageNet (Deng et al., 2009). We first generate positive pairs using augmentation operations such as: cropping, horizontal flipping and color jittering. During training, we set the number of iterations to be 13 and the training epochs in the first stage to be 20 and 5 for the remaining steps. We use stochastic gradient descent as the optimization with a momentum of 0.9 to optimize the model. The learning rate is initialized to 0.1 and changed to 0.01 after 15 epochs. For the clustering stage, the minimum sample parameter for the HDBSCAN (Campello et al., 2013) algorithm is set to 3.

4.4 Comparison with the State of the Art

We compare our proposed CUPR framework with 4 types of state-of-the art unsupervised re-id methods: (1) unsupervised cross-domain re-id models (TJAIDL (Wang et al., 2018), SPGAN (Deng et al.,

Table 1: Comparisons with the state-of-the-art person re-id methods on Market-1501 and DukeMTMC-reID. The 1st and 2nd highest scores are marked by red and blue respectively.

Methods	Labels	Market-1501			DukeMTMC-reID		
		mAP	rank-1	rank-5	mAP	rank-1	rank-5
TJAIDL (Wang et al., 2018)	Cross-domain: labeled source	26.5	58.2	74.8	23.0	44.3	59.6
SPGAN (Deng et al., 2018)		26.9	58.1	76.0	26.4	46.9	62.6
PTGAN (Wei et al., 2018)		15.7	38.6	57.3	13.5	27.4	43.6
HHL (Zhong et al., 2018)		31.4	62.2	78.8	27.2	46.9	61.0
PAUL (Yang et al., 2019)		36.8	66.7	-	35.7	56.1	-
ATNet (Liu et al., 2019)		25.6	55.7	73.2	24.9	45.1	59.5
PUL (Fan et al., 2018)		20.1	44.7	59.1	16.4	30.4	46.04
EUG (Wu et al., 2018)	One labeled image per identity	26.2	55.8	72.3	28.5	48.8	63.4
TAUDL (Li et al., 2018a)	Pure unsupervised: Camera label	41.2	63.7	-	43.5	61.7	-
OIM (Xiao et al., 2017)	Pure unsupervised: No label	14.0	38.0	58.0	11.3	24.5	38.8
BUC (Lin et al., 2019)		29.6	61.9	73.3	22.1	40.4	52.5
TSSL (Wu et al., 2020)		43.3	71.2	-	38.5	62.2	-
Ours		44.6	73.3	84.6	41.4	64.2	76.2

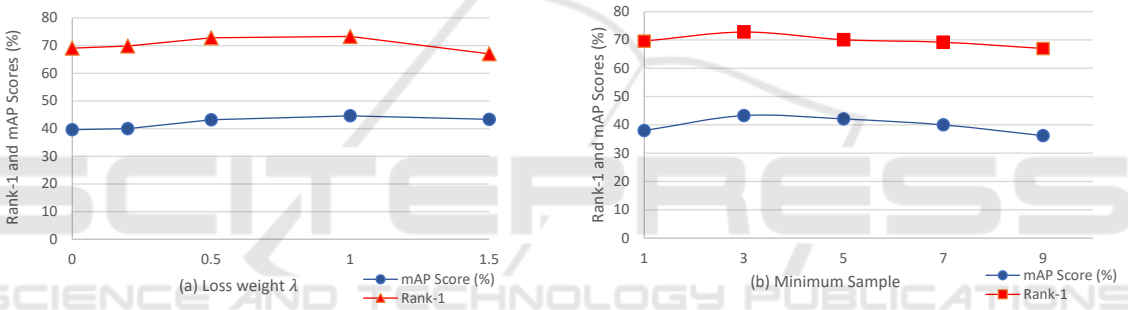

Figure 5: Analysis of hyper-parameters. (a): The impact of the loss weight λ. (b): The impact of the minimum sample S_{min} for each cluster generated by HDBSCAN.

2018) PTGAN (Wei et al., 2018), HHL (Zhong et al., 2018), PAUL (Yang et al., 2019), ATNet (Liu et al., 2019)) and PUL (Fan et al., 2018), (2) one labeled image per identity models (EUG (Wu et al., 2018)), and (3) pure unsupervised re-id models (TAUDL (Li et al., 2018a), OIM (Xiao et al., 2017), BUC (Lin et al., 2019), TSSL (Wu et al., 2020)).

The comparisons are shown in Table 1. On Market-1501, we obtain the best performance among the compared methods with rank-1 = 73.3%, mAP = 44.6%. Therefore, we improved the state of the art TSSL (Wu et al., 2020) by 2.1% and 1.3% respectively, by using a simple learning framework. Compared with OIM (Xiao et al., 2017), we achieve 35.3% more in rank-1 score and 30.6% improvement in mAP. We also compare our method to the state-of-the-art cross-domain methods. Although these methods exploit labelled source data, our approach clearly outperforms the cross-domain state of the art method with 6.6% in rank-1 accuracy and 7.8% in mAP. Similarly, CUPR surpasses all previous methods on DukeMTMC-reID by the best rank-1 accuracy 64.2% and the second best mAP 41.4%. While TAUDL (Li et al., 2018a) performs better in terms of mAP, it assumes extra camera annotation.

Without any human supervising, CUPR outperforms the state-of-the-art methods, which indicates that our method is not only effective in exploiting the unlabeled data, but it is also simple and require no complex architecture.

4.5 Algorithm Analysis

We perform an analysis on three parameters: (1) the loss weight in Eq.(9), (2) the minimum samples parameter for HDBSCAN and (3) the temperature parameter in Eq.(5) to evaluate the parameters sensitivity.

Analysis of the Loss Weight. λ is a hyperparameter used to control the relative importance of the contrastive loss $\mathcal{L}_c$. We sample λ from $\{0, 0.2, 0.5, 1, 1.5\}$

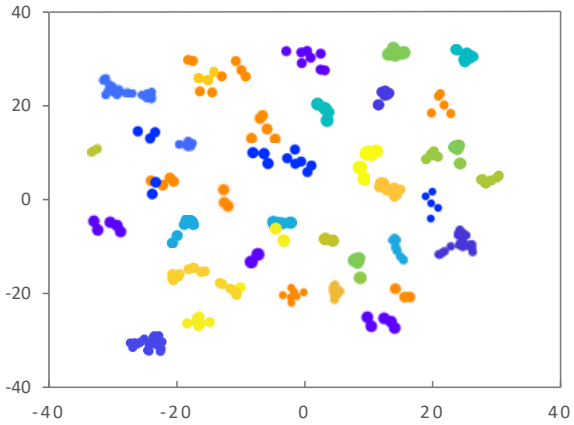

Figure 6: T-SNE visualization of feature distribution of 50 random identities from Market-1501 dataset. Samples with the same color represent the same identity. Samples with different colors grouped together are False Positives. The model groups them by mistake because they look very similar.

to evaluate its impact on the result. As shown in Figure 5.(a), the performance of CUPR increases to 73.3% in rank-1 accuracy and 44.6% in mAP from $\lambda = 0.5$ to $\lambda = 1$, then after start gradually decreasing.

Analysis of the Temperature. τ is a parameter that controls the softness of probability distribution over our clusters. To test its effect, we sample τ from the set $\{0.1, 0.2, 0.5, 1\}$. We observed that the best result is obtained when τ is set to 0.5.

Analysis of the Minimum Sample. In this study, we analyse the effect of the number of minimum samples S_{min} on the Re-ID results. The larger the value of S_{min} we provide, the more conservative our clustering will be. We test the impact of $\{1, 3, 5, 7, 9\}$ minimum samples on the performance of our CUPR framework. As shown in Figure 5.(b), we can see that setting S_{min} to 3 yields to higher accuracy. Meanwhile, larger S_{min} results in declaring lot of points as a noise which negatively impact the accuracy results.

Qualitative Analysis. To further explore the distribution of the learned features, we used T-SNE to visualize the feature embedding of the clusters, by plotting 50 random identities in a 2-dimensional space. As shown in Figure 6, samples of the same identity are grouped together.

5 CONCLUSION

In this paper, we propose a novel contrastive unsupervised learning approach for person re-identification (CUPR). It jointly learns a discriminative representation of unlabeled data and optimizes a CNN model to generate accurate pseudo-labels using clustering. Specifically, the model is trained by considering each sample as an individual cluster. Then, hierarchical clustering algorithm is applied to the feature embedding learned from the model to generate new pseudo-labels. The key component of CUPR is demonstrated by its simple architecture and its ability to exploit unlabeled data. By combining our findings, we improve considerably over previous methods for self supervised, semi-supervised, and cross-domain methods.

REFERENCES

Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15509–15519.

Bian, J.-W., Wu, Y.-H., Zhao, J., Liu, Y., Zhang, L., Cheng, M.-M., and Reid, I. (2019). An evaluation of feature matchers for fundamental matrix estimation. In *British machine vision conference (BMVC)*, volume 2.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., and Jiao, J. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003.

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., and Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774.

Fan, H., Zheng, L., Yan, C., and Yang, Y. (2018). Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–18.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2009). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645.

Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., and Huang, T. (2019). Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8295–8302.

Geng, M., Wang, Y., Xiang, T., and Tian, Y. (2016). Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*.

Gou, M., Karanam, S., Liu, W., Camps, O., and Radke, R. J. (2017). Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–19.

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2019). Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Li, M., Zhu, X., and Gong, S. (2018a). Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision (ECCV)*, pages 737–753.

Li, W., Zhu, X., and Gong, S. (2018b). Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294.

Lin, Y., Dong, X., Zheng, L., Yan, Y., and Yang, Y. (2019). A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8738–8745.

Lisanti, G., Masi, I., Bagdanov, A. D., and Del Bimbo, A. (2014). Person re-identification by iterative re-weighted sparse ranking. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1629–1642.

Liu, J., Zha, Z.-J., Chen, D., Hong, R., and Wang, M. (2019). Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7202–7211.

Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., and Bu, J. (2014). Semi-supervised coupled dictionary learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3557.

Liu, Z., Wang, D., and Lu, H. (2017). Stepwise metric promotion for unsupervised video person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 2429–2438.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., and Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496.

Wang, H., Gong, S., and Xiang, T. (2014). Unsupervised learning of generative topic saliency for person re-identification. In *Proceedings of the British Machine Vision Conference*. BMVA Press.

Wang, J., Zhu, X., Gong, S., and Li, W. (2018). Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2275–2284.

Wei, L., Zhang, S., Gao, W., and Tian, Q. (2018). 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.

Wu, A., Zheng, W.-S., and Lai, J.-H. (2019). Unsupervised person re-identification by camera-aware similarity consistency learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6922–6931.

Wu, G., Zhu, X., and Gong, S. (2020). Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, volume 2.

Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., and Yang, Y. (2018). Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186.

Xiao, T., Li, S., Wang, B., Lin, L., and Wang, X. (2017). Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424.

Yang, Q., Yu, H.-X., Wu, A., and Zheng, W.-S. (2019). Patch-based discriminative feature learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3633–3642.

Ye, M., Lan, X., and Yuen, P. C. (2018). Robust anchor embedding for unsupervised video person re-identification in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 170–186.

Ye, M., Ma, A. J., Zheng, L., Li, J., and Yuen, P. C. (2017). Dynamic label graph matching for unsupervised video re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5142–5150.

Zhao, R., Ouyang, W., and Wang, X. (2013). Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124.

Zhong, Z., Zheng, L., Cao, D., and Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327.

Zhong, Z., Zheng, L., Li, S., and Yang, Y. (2018). Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–188.