Energy Clustering for Unsupervised Person Re-identification

Kaiwei Zeng
College of Computer Science weiliantiankong2011@qq.com

Abstract
Due to the high cost of data annotation in supervised learning for person re-identification (Re-ID) methods, unsupervised learning becomes more attractive in the real world. The Bottom-up Clustering (BUC) approach based on hierarchical clustering serves as one promising unsupervised clustering method. One key factor of BUC is the distance measurement strategy. Ideally, the distance measurement should consider both inter-cluster and intra-cluster distance of all samples. However, BUC uses the minimum distance, only considers a pair of the nearest sample between two clusters and ignores the diversity of other samples in clusters. To solve this problem, we propose to use the energy distance to evaluate both the inter-cluster and intra-cluster distance in hierarchical clustering (E-cluster), and use the sum of squares of deviations (SSD) as a regularization term to further balance the diversity and similarity of energy distance evaluation. We evaluate our method on large scale re-ID datasets, including Market-1501, DukeMTMC-reID and MARS. Extensive experiments show that our method obtains significant improvements over the state-of-the-art unsupervised methods, and even better than some transfer learning methods.

Introduction
Re-ID is a task about whether a person reappears in another camera after being spotted in one, which is widely used in the field of tracking. In recent years, with the development of CNN, supervised learning (Sun et al. 2018; Wang et al. 2018a), which uses the label as a supervisor and lets the model have a clear optimization target, has achieved good performance in Re-ID. But in practical applications, labeling a large amount of data is very expensive, significantly limiting the generalization of supervised learning.

To mitigate the above problem, transfer learning proposed, which only needs labeled source data(Wei et al. 2018; Zhong et al. 2018; 2019; Peng et al. 2016). Unsupervised learning further removes the need of data labels for the source dataset. Traditional unsupervised Re-ID methods include manual features (Farenzena et al. 2010; Lisanti et al. 2014), significant features (Zhao, Ouyang, and Wang 2013) and dictionary learning (Kodirov, Xiang, and Gong 2015; Yan et al. 2018). In recent years, CNN fine-tuning and clustering based Re-ID constitutes the advance in this field. Fan et al. propose PUL (Fan et al. 2018), they 1) learn and extract features of images through CNN fine-tuning; 2) use K-means to cluster different features in each iteration; and 3) regard clusters numbers $k$ as soft labels to classify different images.

However, in PUL, $k$ value needs to be determined in advance and the result of the K-means algorithm is sensitive to the $k$ value. Besides, K-means is a division based clustering method, each iteration needs to iterate over all samples, making the computation huge and difficult to converge. To improve the performance of PUL, Lin et al. propose BUC (Lin et al. 2019). They use hierarchical clustering, which merges a fixed number of clusters and updates the model in each loop iteration. One important factor of the hierarchical clustering method is the distance measurement among clusters. BUC uses the minimum distance. It only calculates the inter-cluster distance, which is the nearest distance of a pair of samples between two clusters, and ignores features of other samples and intra-cluster distance.

However, a good distance measurement should consider both inter-cluster and intra-cluster distance of all samples. Otherwise, some important information will be ignored and result in poor clustering. Besides, Re-ID uses euclidean distance in the calculation. According to the conclusion in (Wang et al. 2017; Ding et al. 2019), it is easy to form elongated clusters which result in poor performance when using the minimum distance. So minimum distance isn’t a good distance measurement.

In order to solve these problems, we propose to use energy distance (Székely and Rizzo 2013; Székely 2003) as the distance measurement. Moreover, as the distribution of samples in datasets is not uniform (shown in Fig.1), we introduce the SSD as a regularization term to measure the dispersion degree within clustering to balance the diversity and similarity of energy distance.

To summarize, the major contributions of this paper are:

- We measure the distance between clusters with energy distance, which can promote more compact clustering.
• We use the SSD as a regularization term, it gives priority to combining single sample clusters and also allows unbalance clustering. In general, it balances the diversity and similarity of clustering.

• Experimental results show that our method achieves the state-of-the-art on Market-1501, DukeMTMC-reID, and MARS in fully unsupervised learning.

Related work
Transfer learning
Transfer learning usually utilizes part of the annotation information which is source dataset labels. Some studies focus on the difference in environments, camera styles in different datasets. Thus, they use GAN to generate new images(Wang et al. 2018b; Qian et al. 2018; Zheng et al. 2019). On the one hand, transferring the background style from the source domain to the target domain can generate new images with labels and decrease differences caused by cameras and environments. On the other hand, it can also expand training datasets, and get better generalization performance. For example, Deng et al. propose SPGAN(Deng et al. 2018) base on CycleGAN(Zhu et al. 2017). The core idea of SPGAN is ID information of images can remain the same before and after transfer learning. To achieve this, they construct an unsupervised self-similarity and domain-dissimilarity relationship to constrain transferring, and transfer the style of training data in the source domain to the target domain. Finally, they use new generated labeled data for training.

Besides, some people inspired by supervised learning. They believe data annotation is equal to a supervisor which can make CNN training has a clear target to optimize and enhance the performance of the model. Thus, during the training, they try to use some methods to tag soft labels for images which without manual annotation labels. Zhong et al. propose ECN(Zhong et al. 2019) base on exemplar-invariance(Wu et al. 2018; Xiao et al. 2017), camera-invariance(Zhong et al. 2018) and neighborhood-invariance(Chen, Zhu, and Gong 2018). They set triple losses about these invariances and use them to expand the distance between different samples and reduce the distance between similar samples. Moreover, they set soft labels for samples, store them in the exemplar memory module(Santoro et al. 2016; Vinyals et al. 2016) and use these soft labels to optimize model in each iteration.

Clustering
Clustering is one of the traditional unsupervised learning methods in machine learning. With the development of CNN, people begin to consciously combine traditional clustering with deep learning(Singh, Gupta, and Efros 2012; Xie, Girshick, and Farhadi 2016). The key of clustering is to compare distances of different features, regard those similar features as the same identities, and gradually reduce the distance of. Fan et al. propose a kind of progressive unsupervised learning combined with K-means in PUL(Fan et al. 2018). They extract features through CNN and divide samples into cluster $C_i$ according to the distance between samples and clustering centers. Finally, they regard $i$ as the soft label of samples. At the beginning, when the model is weak, PUL learns only from a small amount of reliable samples which are close to the cluster centroid in the feature space to avoid falling into local optimal and difficult convergence. As the model becomes more and more powerful in subsequent iterations, more samples will be selected. Finally, the model goes through repeated CNN fine-tuning and K-means clustering until convergence.

Another kind of clustering is hierarchical clustering. Lin et al. propose BUC(Lin et al. 2019) base on it. They extract features by CNN, and then merge a small percent of similar clusters to form a new cluster in each iteration according to the distance between clusters. BUC compares the mini-
Hierarchical clustering
A given training set of $N$ images $X = \{x_1, x_2, \ldots, x_N\}$, we have manual annotation about $n$ identities, label $= \{y_1, y_2, \ldots, y_n\}$ in supervised learning. Therefore, we can directly take the label as the supervision and use CNN for optimization. In general, we usually use softmax and cross entropy to measure the confidence of a predict label:

$$P(y_i|x) = \text{softmax}(W_i^T x) = \frac{\exp(W_i^T x)}{\sum_{j=1}^{n} \exp(W_j^T x)}$$  \hspace{1cm} (1)

where $y_i$ is the predicted label, $P(y_i|x)$ is the predicted probability of $x$ belonging to $y_i$, $W$ is the weight of model. We can learn $y = f(W, x)$ directly from CNN in supervised learning. But in unsupervised learning, because there is no any manual annotation like $y_i$, we need to learn the feature embedding function $\phi(\theta; x)$, where $\theta$ is the weight of CNN. For the query set $\{x_i^q\}_{i=1}^{N_q}$ and the gallery set $\{x_i^g\}_{i=1}^{N_g}$, we need to compare the similarity of images: $d(x_i^q, x_i^g) = ||\phi(x_i^q; \theta) - \phi(x_i^g; \theta)||$ to determine the image identity according to the distance between features during the evaluation. Finally, we divide similar images into the cluster $C_i$ and regard $i$ as the soft label of images in cluster $C_i$. Our network structure is shown in Fig.2, which mainly includes two parts. One is to extract sample features by CNN fine-tuning; the other is to merge clusters by hierarchical clustering and repeatedly iterate to convergence. We define the probability that image $x$ belongs to the $i$-th cluster as:

$$p(c|x, V) = \frac{\exp(V_i^T v/\tau)}{\sum_{j=1}^{n} \exp(V_j^T v/\tau)}$$  \hspace{1cm} (2)

where $V$ is the lookup table which contains all features, $V_i$ is the $i$-th cluster lookup table, contains all samples in $i$-th cluster, $n$ is the cluster number in current iteration, $v$ is the $L_2$ normalized feature obtained from CNN. $\tau$ is a temperature parameter(Hinton, Vinyals, and Dean 2014) that
controls the softness of probability distribution. According to Xiao et al. (2017), we set $\tau = 0.1$.

**Energy distance**

In hierarchical clustering, how to measure the distance between clusters determines the performance of the model. Energy distance is the distance between statistical observations in metric spaces. This concept derives from Newton’s gravitational potential energy. Energy distance views the sample as an object subject to statistical potential energy. According to Székely and Rizzo (2013), they measure the distance by:

$$2 \int_{-\infty}^{\infty} (F(x) - G(x))^2 \, dx$$

(3)

where $F(\cdot), G(\cdot)$ are cumulative distribution functions (cdf) of a random variable. For $X, Y$ are independent random variables with cdf $F$ and $G$:

$$2 \int_{-\infty}^{\infty} (F(x) - G(x))^2 \, dx$$

$$= 2E|X - Y| - E|X - X'| - E|Y - Y'|$$

(4)

where $X'$ is an independent and identically distributed (iid) copy of $X, Y'$ an iid copy of $Y, E(\cdot)$ is the mathematical expectation of samples. When it is extended to the rotation invariant high dimensional space, the energy distance is:

$$\mathcal{E}(X, Y) = 2E|X - Y| - E|X - X'| - E|Y - Y'|$$

(5)

where $X, Y \in \mathbb{R}^d$ are two iid sample space, $d$ is the dimension of the sample space. Székely et al. prove that Eq.(5) is nonnegative and equals zero if and only if $X$ and $Y$ are iid in Székely and Rizzo (2005b). When it is extended to multi-sample energy distance:

$$\mathcal{E}(X, Y) = \frac{2}{nm} \sum_{i,j=1}^{n,m} ||x_i - y_j||_2 - \frac{1}{n^2} \sum_{i,j=1}^{n} ||x_i - x_j||_2 - \frac{1}{m^2} \sum_{i,j=1}^{m} ||y_i - y_j||_2$$

(6)

where $X = \{x_1, x_2, \cdots, x_n\}, Y = \{y_1, y_2, \cdots, y_m\}$ and $n,m$ are the number of samples in $X, Y, ||\cdot||_2$ means the euclidean distance between samples. The ability of energy distance to separate and identify clusters with equal or near equal centers is an important practical advantage over geometric clustering center methods, such as centroid, minimum, maximum and Ward’s minimum variance methods. In Székely and Rizzo (2005a), the simulation results show that energy clustering can effectively recover the underlying hierarchical structure under different scenarios, including high dimensional data and data with different scale attributes, while maintaining the advantages of separating spherical clustering.

**Regularization term**

In the process of clustering, we need to consider the compact degree of each cluster as well as the distance of inter-cluster. The sum of squares of deviations (SSD) is usually used to measure the sum deviation between each sample and the mean. In other words, it equals to variance*$n$. For sample space $X = \{x_1, x_2, \cdots, x_n\}$, the mean is $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Figure 3: We use ResNet-50 to extract features for all images, use E-cluster to merge clusters, and finally update soft labels as the input of the next iteration. (a) - (b) describe the process of using the SSD as the regularization term to balance the similarity and diversity of E-cluster. (a) For the single-sample cluster, it will be merged preferentially because its SSD is zero. So the yellow and the gray sample will be merged. (b) The SSD measures the dispersion of samples in clusters and it isn’t strictly linearly dependent on the number of samples. As a result, it doesn’t exclude the formation of super clusters. As shown in the figure, the brown cluster will merge the green cluster, instead of the red cluster, because it has a lower SSD.
the SSD is:

$$SSD(X) = \sum_{i=1}^{n} (x_i - \bar{X})^2$$ (7)

At the beginning, SSD in the single-sample cluster is zero, so we can combine the single-sample cluster first. Besides, it doesn’t exclude the form of super clusters because SSD considers both sample numbers and the dispersion of samples in clusters. In order to balance the SSD and energy distance, we introduce the trade-off parameter $\lambda$, and finally our merge rule is defined as:

$$D_{dest}(X, Y) = E(X, Y) + \lambda SSD(X)$$ (8)

**Update and merge**

As shown in the algorithm, we will regard $N$ samples as $N$ different identities at the beginning and use CNN extract features. We believe that images with the same identities will close to each other in high dimensional space because of their similarity. Therefore, we measure the distance of each cluster according to the energy distance and merge a part of clusters in each step. We set two hyperparameters, one is $\lambda$, which is used to balance the energy distance and regularization term, and the other is $mp$, which is used to control the clustering merging speed. Besides, $n$ represents the current cluster number, and $m = N \times mp$ represents the number of clustering merged in each step. That is, $m$ pairs of clusters with the nearest distance will be merged into one respectively. We iterate the model until we observe a performance drop on the validation set.

### Experimental Results

**Datasets**

**Market-1501** Market-1501(Zheng et al. 2015) is consist of 1,501 identities observed under 6 camera viewpoints. Each pedestrian is captured by at least two cameras and may have multiple images in a single camera. The training set contains 751 identities about 12,936 images, and the test set contains 750 identities about 19,732 images.

**DukeMTMC-reID** DukeMTMC(Ristani et al. 2016) contains 85 minutes of high-resolution video from eight different cameras. DukeMTMC-reID(Zheng, Zheng, and Yang 2017) is a subset of DukeMTMC. It contains 16,522 images for training, 2,228 images for query, and 17,661 images for gallery

**Mars** Mars(Zheng et al. 2016) is an extension of Market-1501 dataset, the acquisition method is same as Market-1501. It contains 17,503 video traces of 1,261 identities, 625 for training and 636 for testing.

**Experimental Settings**

**Training** For image Re-ID dataset Market-1501 and DukeMTMC-reID, we only use pictures that remove labels to train CNN. For video Re-ID dataset MARS, we only use tracklets of identities, and each tracklet is regarded as an individual. Note that our method is fully unsupervised learning because we don’t use any manual annotation.

**Algorithm 1 E-cluster Algorithm**

**Require:**

- Input $X = \{x_i\}_{i=1}^{N}$
- Merging percent $m \in (0, 1)$
- Hyperparameter $\lambda$
- Initial model $\phi(\cdot; \theta_0)$

**Ensure:**

- Best model $\phi(\cdot; \theta)$

1: Initialize:soft labels $Y = \{y_i = i\}_{i=1}^{N}$, cluster number $n = N$, merging number $m = n \times mp$
2: while $n > m$ do
3: \hspace{1em} Train model with $X, Y$;
4: \hspace{1em} Extract feature and update lookup table $V$;
5: \hspace{1em} Calculate energy distance and variance between clusters in $V$, according to Eq.(8)
6: \hspace{1em} Select clusters to merge: $n = n - m$
7: \hspace{1em} Update $Y$ with new soft labels:
8: \hspace{2em} $Y = \{y_i = j, \forall x_i \in C_j\}_{i=1}^{N}$
9: \hspace{2em} if $m \bar{AP}_i > m \bar{AP}_{best}$ then
10: \hspace{3em} $m \bar{AP}_{best} = m \bar{AP}_i$
11: \hspace{3em} Best model=$\phi(x; \theta_i)$
12: \hspace{1em} end if
13: end while

**Evaluation** we use the mean average precision (mAP) and the rank-$k$ accuracy to evaluate the performance of the model. The mAP is calculated according to the precision-recall curve, reflecting the overall accuracy and recall rate. Rank-$k$ emphasizes the accuracy of retrieval, it means the query picture have the match in the top-$k$ list.

**Experimental details** In the experiment, Resnet-50(He et al. 2016) is used as the backbone network to extract features. We also adopt pre-training weights on ImageNet, remove the last classification layer and add a FC layers behind it as embedding features. Without special instructions, we usually keep embedding dimension as 2048-d. During the training of CNN, we set the number of training epochs in the first stage to be 20 and in the following stage to be 2 for fine-tuning. Beside, we set batch size to be 16, and dropout rate to be 0.5. All the networks are trained using stochastic gradient descent (SGD) with the momentum of 0.9. The learning rate is initialized to 0.1 and decreased to 0.01 after 15 epoches. For the clustering stage, the merging percent $m$ is set to be 0.05 and $\lambda$ to be 0.9.

**Comparison with state-of-the-art**

**Comparison with hierarchical clustering** We compare our method with BUC and DBC in Table1. On Market-1501, BUC obtain rank-1 = 66.2%, mAP = 38.3%, DBC get rank-1 = 69.2%, mAP = 41.3%. E-cluster obtain the best result with rank-1 = 70.7%, mAP = 43.0%. Minimum distance only considerate one pair of sample and ignore others. DBC focus on average pairwise distance about all pairs of samples, but ignore a special situation. Energy distance has an advantage that if and only if $X, Y$ are iid, the distance between $X, Y$ should be zero. We can see E-cluster get significant
Table 1: The comparison of E-cluster with the BUC and DBC, our regularization method is SSD, DBC is intra-cluster discrete degree, BUC is the sample number in each cluster.

| Method                      | Market-1501 | DukeMTMC-reID | MARS |
|-----------------------------|-------------|---------------|------|
|                            | rank-1 | mAP | rank-1 | mAP | rank-1 | mAP  |
| BUC without regularizer     | 62.9   | 33.8 | 41.3   | 22.5 | 55.5   | 31.9 |
| BUC with regularizer        | 66.2   | 38.3 | 47.4   | 27.5 | 61.1   | 38.0 |
| DBC with regularizer        | 66.2   | 38.7 | 48.2   | 27.5 | 59.8   | 37.2 |
| DBC without regularizer     | 69.2   | 41.3 | 51.5   | 30.0 | 64.3   | 43.8 |
| E-cluster without regularizer| 68.5   | 40.7 | 50.7   | 28.6 | 61.1   | 38.3 |
| E-cluster with regularizer  | 70.7   | 43.0 | 52.7   | 31.4 | 65.1   | 44.2 |

Table 2: We evaluate E-cluster and compare to recent methods on Market-1501 and DukeMTMC-reID, the label column lists the type of supervision used by the method. "Transfer" means it uses an external dataset with annotations, "OneEx" means only one image per identity is labeled, "None" denotes no additional information is used. * denotes results are reproduced by (Lin et al. 2019).

| Methods                      | Labels | Market-1501 | DukeMTMC-reID |
|------------------------------|--------|-------------|---------------|
|                              |        | rank-1 | rank-2 | rank-10 | mAP | rank-1 | rank-2 | rank-10 | mAP |
| BOW(Zheng et al. 2015)       | None   | 35.8   | 52.4   | 60.3   | 14.8 | 17.1   | 28.8   | 34.9   | 8.3  |
| OIM*(Xiao et al. 2017)       | None   | 38.0   | 58.0   | 66.3   | 14.0 | 24.5   | 38.8   | 46.0   | 11.3 |
| UMDL(Peng et al. 2016)       | Transfer | 34.5  | 52.6   | 59.6   | 12.4 | 18.5   | 31.4   | 37.6   | 7.3  |
| PUL(Fan et al. 2018)         | Transfer | 44.7  | 59.1   | 65.6   | 20.1 | 30.4   | 46.4   | 50.7   | 16.4 |
| EUG(Wu et al. 2019)          | OneEX  | 49.8   | 66.4   | 72.7   | 22.5 | 45.2   | 59.2   | 63.4   | 24.5 |
| SPGAN(Deng et al. 2018)      | Transfer | 58.1  | 76.0   | 82.7   | 26.7 | 49.6   | 62.6   | 68.5   | 26.4 |
| TJ-AIDL(Wang et al. 2018b)   | Transfer | 58.2  | -      | -      | 26.5 | 44.3   | -      | -      | 23.0 |
| BUC(Lin et al. 2019)         | None   | 66.2   | 79.6   | 84.5   | 38.3 | 47.4   | 62.6   | 68.4   | 27.5 |
| DBC(Ding et al. 2019)        | None   | 69.2   | 83.0   | 87.8   | 41.3 | 51.5   | 64.6   | 70.1   | 30.0 |
| E-cluster without regularizer| None   | 68.5   | 83.2   | 87.4   | 40.7 | 50.7   | 63.7   | 69.7   | 28.6 |
| E-cluster with regularizer   | None   | 70.2   | 84.1   | 88.6   | 42.8 | 52.7   | 66.1   | 70.6   | 31.4 |

improvement and it proves energy distance do better to performance than minimum distance and DBC on unsupervised re-ID learning. Besides, it can also illustrate SSD is an appropriate standard to balance diversity and similarity in clusters.

Image-based Person Re-identification Table 2 reports the result of state-of-the-art unsupervised Re-ID methods on image-based person Re-ID datasets. On Market-1501, we achieve the best performance with rank-1 = 69.2%, mAP = 41.3% among all fully unsupervised methods, BOW, OIM, BUC and DBC. Similarly, we also achieve 52.7% in rank-1 and 31.4% in mAP on DukeMTMC-reID and beyond other unsupervised methods. Besides, we also compare E-cluster with some transfer learning method. Although these methods make use of partially manually annotated data as external supervision compared to fully unsupervised learning, our e-cluster still performs better than some domain transfer methods.

Video-based Person Re-identification Table 3 reports the result of state-of-the-art unsupervised Re-ID methods on video-based person Re-ID datasets. On Mars, we achieved rank-1=65.1%, mAP=44.2%. This illustrates better generalization ability of our method on different data distributions. We also compare our method to the state-of-the-art methods in the video-based one-example setting. According to (Wu et al. 2018), OneEX uses annotating labels from a tracklet for each person which are not fully unsupervised. E-cluster still get the best performance and prove our method make more efficient use of unlabeled data.

Discussion

Effective of Energy distance

In order to evaluate the effectiveness of energy distance, we compare E-cluster with BUC and DBC. The common basis of three methods is hierarchical clustering, the difference is the distance measurement used in the merge stage. To be fair, we don’t use regularization term in comparison. Fig. 4 shows the result on Market-1501. We compare the entire training iteration process of the three methods by taking mAP as the indicator. We still set the merge percent to 0.05, so there are 20 iterations in total. At first, there isn’t much difference among the three methods. In the middle stage of training, performance differences gradually increase and E-cluster gets the best. In the end, we observe obvious performance degradation in all methods but E-cluster still keeps the best.

Of course, this also shows a problem. No matter what distance measurement method is selected, the algorithm per-
Table 3: Results on video-based Re-ID datasets. "OneEx" means only one image in per identity is labeled, "None" denotes no additional information is used. * denotes that the results are reproduced by (Lin et al. 2019).

| Methods                  | Labels | MARS |
|--------------------------|--------|------|
|                          | rank-1 | rank-5 | rank-10 | mAP  |
| OIM* (Xiao et al. 2017)  | None   | 33.7  | 48.1   | 54.8 | 13.5 |
| DGM+IDE (Ye et al. 2019) | OneEx  | 36.8  | 54.0   | -    | 16.8 |
| Stepwise (Liu, Wang, and Lu 2017) | OneEx   | 41.2  | 55.5   | -    | 19.6 |
| RACE (Ye, Lan, and Yuen 2018) | OneEx  | 43.2  | 57.1   | 62.1 | 24.5 |
| DAL (Chen, Zhu, and Gong 2018) | OneEx  | 49.3  | 65.9   | 72.2 | 23.0 |
| BUC (Lin et al. 2019)    | None   | 61.1  | 75.1   | 80.0 | 38.0 |
| EUG (Zhong et al. 2019)  | OneEx  | 62.6  | 74.9   | -    | 42.4 |
| DBC (Ding et al. 2019)   | None   | 64.3  | 79.2   | 85.1 | 43.8 |
| E-cluster                | None   | 65.1  | 78.8   | 85.0 | 44.2 |

Figure 4: Comparison about results with clustering stages on Market-1501 without regularization term. Compared to BUC and DBC, E-cluster have a better robustness in later stages and get the best performance.

Performance will decline in the later stage of hierarchical clustering. We explain that there are a large number of small clusters in the early stage, even if there are some errors, we still get steady improvement in performance. But at the end of the algorithm, there is a lot of big clusters, merge errors in early stages have a superposition effect. Therefore, at this point, the wrong merging will have a great influence and lead to a decline in performance. An effective solution is to propose a more effective distance measurement method to reduce the errors of the algorithm when merging small clusters in the early stage. This is also the next improvement direction of hierarchical clustering.

**Trade-off parameter analysis**

The regularization term balances the energy distance and intra-clusters variance. In order to assessment the influence of regularization coefficient on the final result and get the best performance, we compare the performance at different $\lambda$ values on market-1501 dataset. The result is reported in Fig.5. When $\lambda$ increases from 0 to 0.9, the model performance reaches its peak, and then further increasing $\lambda$ will lead to obvious performance decline. This is very intuitive because too large $\lambda$ will lead to high SSD weight in Eq.(8). Further it will mask the superior properties of the energy distance and mAP will even be lower than the original E-cluster.

**Conclusions**

In this paper, we emphasize the importance of distance measurement for the performance in hierarchical clustering model. We analyze the deficiency of minimum distance and the advantage of energy distance. Base on these, we propose a new clustering connection standard, which combines the energy distance with the variance. The energy distance makes clusters more compact and the SSD keeps the internal dispersion smaller. We balance the diversity and similarity in clusters and extensive experiments show that our method has better performance.

**References**

Chen, Y.; Zhu, X.; and Gong, S. 2018. Deep association learning for unsupervised video person re-identification. In *BMVC*. 
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 CVPR.

Ding, G.; Khan, S.; Tang, Z.; Zhang, J.; and Porikli, F. 2019. Towards better validity: Dispersion based clustering for unsupervised person re-identification. arXiv:1906.01308.

Fan, H.; Zheng, L.; Yan, C.; and Yang, Y. 2018. Unsupervised person re-identification: Clustering and fine-tuning. TOMM.

Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; and Cristani, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In ICCV.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In CVPR.

Hinton, G.; Vinyals, O.; and Dean, J. 2014. Distilling the knowledge in a neural network. In NIPS-W.

Kodirov, E.; Xiang, T.; and Gong, S. 2015. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In BMVC.

Lin, Y.; Dong, X.; Zheng, L.; Yan, Y.; and Yang, Y. 2019. A bottom-up clustering approach to unsupervised person re-identification. In AAAI.

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.

Liu, Z.; Wang, D.; and Lu, H. 2017. Stepwise metric promotion for unsupervised video person re-identification. In ICCV.

Peng, P.; Xiang, T.; Wang, Y.; Pontil, M.; Gong, S.; Huang, T.; and Tian, Y. 2016. Unsupervised cross-dataset transfer learning for person re-identification. In CVPR.

Qian, X.; Fu, Y.; Xiang, T.; Wang, W.; Qiu, J.; Wu, Y.; Jiang, Y.-G.; and Xue, X. 2018. Pose-normalized image generation for person re-identification. In ECCV.

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 ECCV. Springer.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. Meta-learning with memory-augmented neural networks. In ICML.

Singh, S.; Gupta, A.; and Efros, A. A. 2012. Unsupervised discovery of mid-level discriminative patches. In ECCV. 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 ECCV.

Szekely, G. J., and Rizzo, M. L. 2005a. Hierarchical clustering via joint between-within distances: Extending ward’s minimum variance method. Journal of classification.

Szekely, G. J., and Rizzo, M. L. 2005b. A new test for multivariate normality. Journal of Multivariate Analysis.

Szekely, G. J., and Rizzo, M. L. 2013. Energy statistics: A class of statistics based on distances. Journal of statistical planning and inference.

Szekely, G. J. 2003. E-statistics: The energy of statistical samples. Bowling Green State University, Department of Mathematics and Statistics Technical Report.

Szekely, G. J. 2012. Contests in higher mathematics: Miklós Schweizer Competitions 1962–1991. Springer Science.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In NIPS.

Wang, F.; Xiang, X.; Cheng, J.; and Yuille, A. L. 2017. Normface: 12 hypersphere embedding for face verification. In ACM Multimedia. ACM.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018a. Learning discriminative features with multiple granularities for person re-identification. In ACM Multimedia. ACM.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2018b. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In CVPR.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In CVPR.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In CVPR.

Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Bian, W.; and Yang, Y. 2019. Progressive learning for person re-identification with one example. IEEE Transactions on Image Processing.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In CVPR.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In ICML.

Yan, C.; Luo, M.; Liu, W.; and Zheng, Q. 2018. Robust dictionary learning with graph regularization for unsupervised person re-identification. Multimedia Tools and Applications.

Ye, M.; Li, J.; Ma, A. J.; Zheng, L.; and Yuen, P. C. 2019. Dynamic graph co-matching for unsupervised video-based person re-identification. IEEE Transactions on Image Processing.

Ye, M.; Lan, X.; and Yuen, P. C. 2018. Robust anchor embedding for unsupervised video person re-identification in the wild. In ECCV.

Zhao, R.; Ouyang, W.; and Wang, X. 2013. Unsupervised salience learning for person re-identification. In CVPR.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In ICCV.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In ECCV. Springer.

Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; and Kautz, J. 2019. Joint discriminative and generative learning for person re-identification. In CVPR.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In ICCV.

Zhong, Z.; Zheng, L.; Li, S.; and Yang, Y. 2018. Generalizing a person retrieval model hetero-and homogeneously. In ECCV.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In CVPR.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In ICCV.