Transductive Centroid Projection for Semi-supervised Large-scale Recognition

Yu Liu[0000−0001−5812−1137] [1,2], Guanglu Song[2], Jing Shao[2], Xiao Jin[2], and Xiaogang Wang[1,2]

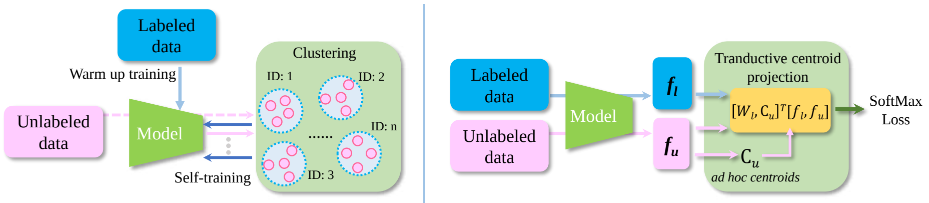
[1] The Chinese University of Hong Kong, Shatin, Hong Kong
{yuliu,xgwang}@ee.cuhk.edu.hk
[2] Sensetime Group Limited, Beijing 100084, China
{songguanglu,jinxiao,shaojing}@sensetime.com

Abstract. Conventional deep semi-supervised learning methods, such as recursive clustering and training process, suffer from cumulative error and high computational complexity when collaborating with Convolutional Neural Networks. To this end, we design a simple but effective learning mechanism that merely substitutes the last fully-connected layer with the proposed Transductive Centroid Projection (TCP) module. It is inspired by the observation of the weights in the final classification layer (called *anchors*) converge to the central direction of each class in hyperspace. Specifically, we design the TCP module by dynamically adding an *ad hoc anchor* for each cluster in one mini-batch. It essentially reduces the probability of the inter-class conflict and enables the unlabelled data functioning as labelled data. We inspect its effectiveness with elaborate ablation study on seven public face/person classification benchmarks. Without any bells and whistles, TCP can achieve significant performance gains over most state-of-the-art methods in both fully-supervised and semi-supervised manners.

Keywords: Person Re-ID · Face Recognition · Deep Semi-supervised Learning

1 Introduction

The explosion of the Convolutional Neural Networks (CNNs) brings a remarkable evolution in the field of image understanding, especially some real-world tasks such as face recognition [1–5] and person re-identification (Re-ID)[6–11]. Much of this progress was sparked by the creation of large-scale datasets as well as the new and robust learning strategies for feature learning. For instance, MS-Celeb-1M [12] and MARS [13] provide more than 10-million face images and 1-million pedestrian images respectively with rough annotation. Moreover, in the industrial environment, it may take only a few weeks to collect billions of face/pedestrian gallery from a city-level surveillance system. But it is hard to label such billion-level data. Utilizing these large-scale unlabelled data to benefit the classification tasks remains non-trivial.



(a) Unstable Self-training strategy for semi-supervised learning (b) TCP enable training unlabelled data with inornate classification pipeline

Fig. 1. A comparison between (a) self-training process with recursive clustering-finetuning (b) un/semi-supervised learning with transductive centroid projection

Most of recent unsupervised or semi-supervised learning approaches for face recognition or Re-ID [14–20] are based on self-training, *i.e.* the model clusters the training data and then the clustered results are used to fine-tune the model iteratively until converges, as shown in Fig. 1(a). The typical downsides in this process lie in two aspects. First, the recursive training framework is time-consuming. And second, since the clustering algorithms used in such approaches always generate ID-clusters with high precision scores but somewhat low recall score, that guarantee the clean clusters without inner errors, it may cause *inter-class conflict*, *i.e.* instances belonging to one identity are divided into different clusters, which hampers the fine-tuning stage. To this end, a question arises: how to utilize unlabelled data in a stable training process, such as a CNN modle with softmax classification loss function, without any recursion and avoid the inter-class conflict?

In this study, we design a novel Transductive Centroid Projection layer to efficiently incorporate the training of the unlabelled clusters accompanied by the learning of the labelled samples, and can be readily extended to an unsupervised manner by setting the labelled data to $\varnothing$.

It is enlightened from the latent space learned by the common used Softmax loss. In deep neural network, each column in the projection matrix $\mathbf{W}$ of the final fully-connected layer indicates the *normal direction* of the decision hyperplane. We call each column as *anchor* in this paper. For a labelled data, the *anchor* of its class already exists in $\mathbf{W}$, and thus we can train the network by maximizing the inner product of its feature and its anchor. However, the unlabelled data doesn't even have a class, so it cannot directly provide the decision hyperplane. To utilize unlabelled samples with conventional deep classification network, we need to find a way to simulate the their *anchors*.

Motivated by the observation that the *anchor* approximates the centroid direction as shown in Fig. 2, the transductive centroid projection layer could dynamically estimate the class centroids for the unlabelled clusters in each mini-batch, and treat them as the new anchors for unlabelled data which are then absorbed to the projection matrix so as to enable classification for both labelled and unlabelled data. As visualized in Fig. 1(b), the projection matrix $\mathbf{W}$ of the classification layer in original CNN is replaced by the joint matrix of $\mathbf{W}$ and *ad hoc* centroids $\mathbf{C}$. In this manner, labelled data and unlabelled data function the same during training. As analyzed in Sec. 3.3, since the *ad hoc* centroids in each mini-batch is much fewer than the total cluster number, the inter-class conflict ratio is naturally low and can hardly influence the training process.

Comprehensive evaluations have been conducted in this paper to compare with some popular semi-supervised methods and some loss functions in metric learning. The proposed transductive centroid projection has a superior performance on stabilizing unsupervised/semi-supervised and optimizing the learned feature representation.

To sum up, the contribution of this paper is threefold:
1) *Observation interpretation* - We investigate the observation that the directions of anchor (*i.e.* weight $\mathbf{w}_n$) gradually coincides with the centroid as model converges, both theoretically and empirically.
2) *A novel Transductive Centroid Projection layer* - Based on the observation above, we propose an innovative un/semi-supervised learning mechanism to wisely integrate the unlabelled data into the recognition to boost its discriminative ability by introducing a new layer named as Transductive Centroid Projection (TCP). Without any iterative processing like self-training and label propagation, the proposed TCP can be simply trained and steadily embedded into arbitrary CNN structure with any classification loss.
3) *Superior performance on face recognition and ReID benchmarks* - We apply TCP to the task of face recognition and person re-identification, and conduct extensive evaluations to thoroughly examine its superiority to both semi-supervised learning and supervised learning approaches.

1.1 Related works

Semi-supervised learning. An effective way for deep semi-supervised learning is the label propagation with self-training [21] by trusting the predicted label from the model trained on labeled data or clustered by clustering model [22–25], for close set or open set respectively. It will hamper the model convergence if the threshold is not precisely set. Other methods like Generative models [26], semi-supervised Support Vector Machines [27] and some graph-based semi-supervised learning methods [28] hold clear mathematical framework but are hard to be incorporated with deep learning methods.
Semi-supervised face/person recognition. In [16], a couple dictionaries are jointly learned from both labelled and unlabelled data. LSRO [8] adopts GAN [29] to generate person patches to normalize data distribution and propose a loss named LSRO to supervise the generated patches. Some works [19, 18] adopt local metric loss functions (*e.g.* triplet loss [2]) to avoid the inter-class conflict. These methods with local optimization function, however, are usually unstable and hard to converge, especially for large-scale data. Some other methods [19] adopt softmax loss to optimize global classes and suffer from the inter-class conflict. Most of these methods focus on transfer learning, self-training and data distribution normalization. In this work, we mainly pay attention to a basic question, namely how to wisely train a simple CNN model by fully leveraging both labelled and unlabelled data, without self-training or transfer learning.

Table 1. Experimental settings on three tasks with different data scales to validate the observation

Task	#Class	Backbone	#Feature Dim.	Feature Space
MNIST	10	LeNet [30]	2	Fig. 2(a)
CIFAR-100	100	ResNet-18 [31]	128	Fig. 2(b)
MS1M-100K	100,000	Inception-ResNet [32]	128	Fig. 2(c)

2 Observation inside the Softmax Classifier

In a typical straightforward CNN, let $\mathbf{f} \in \mathbb{R}^D$ denote the feature vector of one sample generated by prior layers, where D is the feature dimension. The linear activation $\mathbf{y} \in \mathbb{R}^N$ referring to N class labels is therefore accompanied with the weight $\mathbf{W} \in \mathbb{R}^{D\times N}$ and bias $\mathbf{b} \in \mathbb{R}^N$,

$$\mathbf{y} = \mathbf{W}^T\mathbf{f} + \mathbf{b}. \tag{1}$$

In this work we degenerate this classifier layer from affine to linear projection by setting the bias term $\mathbf{b} \equiv \mathbf{0}$. Supervised by softmax loss and optimized by SGD, we can usually observe the following phenomenon: The anchor $\mathbf{w}_i = \mathbf{W}_{[i]} \in \mathbb{R}^D$ for class i points to the direction of the data centroid of class i, when the model has successfully converged. We first show this observation in three toy examples from a low-dimensional space to a high one. Then we try to interpret it by gradient view.

2.1 Toy Examples

To investigate the aforementioned observation from small-scale to large-scale tasks and from low dimensional to high dimensional latent space, we empirically analyze three tasks with different data scales, feature dimention and network structure, *i.e.* character classification on MNIST [33] with 10 classes, object classification on CIFAR-100 [34] with 100 classes, and face recognition on M-S1M [35] with $100,000$ classes[3]. Table 1 records the detailed settings for these experiments. To each task, there are two `FC` layers after its backbone structure, in which `FC1` learns an internal feature vector $\mathbf{f}$ and `FC2` acts as the projection onto the class space. All tasks employ the softmax loss. Fig. 2 depicts the feature spaces extracted from different datasets, in which the 2-D features in MNIST are directly plotted and the 128-D features in CIFAR-100 and MS1M are compressed by Barnes-Hut t-SNE [36].

MNIST – Fig. 2(a) describes the feature visualization in three stages: 0, 2 and 10 epochs. We set the feature dimension $D = 2$ for $\mathbf{f}$ so as to explore the distribution in low dimensional case. The training of this model progressively increases the congregation between features in each class and inter-discrepancy between classes. We pick four classes and show their directions $\mathbf{W}_{[n]}$ from the

[3] The original MS1M dataset has one million face identities with several noises samples. Here we only take the first $100,000$ identities for the convenience of illustration.

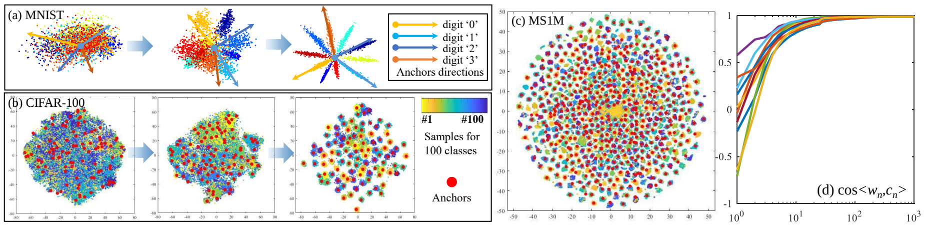


Fig. 2. Visualization of feature spaces on different tasks, *i.e.* (a) MNIST, (b) CIFAR-100 and (c) MS1M, where the features of CIFAR-100 and MS1M are visualized by Barnes-Hut t-SNE [36], and (d) depicts the evolution of cosine distance between anchor direction and class centroid with respect to the training iteration on MNIST

projection matrix $\mathbf{W}$, named as *anchor*. All anchors have random directions at the initial stage of training, and they gradually move towards the direction of their respective centroids.

CIFAR-100 & MS1M – To examine this observation in a much larger data scale and higher dimension case, we further apply CIFAR-100 and MS1M for an ample demonstration. Different from MNIST, the feature dimension for $\mathbf{f}$ is $D = 128$ and t-SNE is used for dimensionality reduction without losing cosine metric. Similar to the phenomenon as observed in MNIST, features in each class tend to be progressively clustered together while features from different classes own more distinct margins in between. Meanwhile the anchors marked by red dots almost locate around its corresponding class centroids. The anchors of a well trained MS1M model also co-locate with the class centroids.

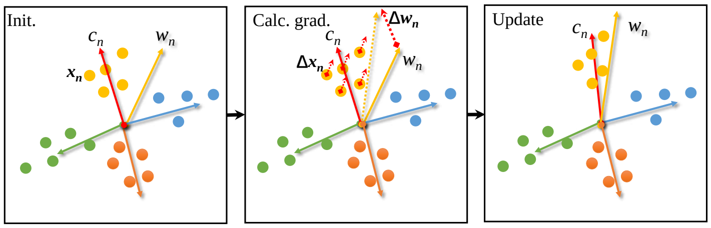
In addition, for a quantified assessment, we compute the cosine similarity $\mathcal{C}(\mathbf{w}_n, \mathbf{c}_n)$ between the anchor $\mathbf{w}_n = \mathbf{W}_{[n]}$ and the class centroid $\mathbf{c}_n$ for the n^{th} class out of 10 classes in total on MNIST. Fig. 2(d) exhibits $\mathcal{C}(\mathbf{w}_n, \mathbf{c}_n)$ with respect to the training iterations. Almost all classes converge to a distance of 1 within one epoch, *i.e.* the direction of the anchor shifts to the same direction of the class centroid.

To conclude, the anchor direction $\mathbf{W}_{[n]}$ is always consistent with the direction of the corresponding class centroid over different dataset scales with various lengths of the feature dimension in $\mathbf{f}$.

2.2 Investigate in Gradients

We investigate the reason why the directions of anchor and centroid will be gradually consistent, from the perspective of gradient descent in the training procedure. Considering the input of linear projection $\mathbf{f}$ which belongs to the n-th chass and the output $\mathbf{y} = \mathbf{W}^T\mathbf{f}$, the softmax probability of $\mathbf{f}$ belongs to n-th chass can be calculated by:

$$p_n = softmax(y) = \frac{\exp(\mathbf{y}_n)}{\sum_{i=1}^{N} \exp(\mathbf{y}_i)} \tag{2}$$



(a) Forward $\boldsymbol{x}$, get $\boldsymbol{c_n}$ (b) BP, get $\boldsymbol{\Delta x_n}$ and $\boldsymbol{\Delta w_n}$ (c) Update $\boldsymbol{w_n}$ and $\boldsymbol{x_n}$ and re-calculate $\boldsymbol{c_n}$

Fig. 3. The evolution of the anchor w_n and features x_n for class n within one iteration. After this iteration, the directions between anchor w_n and centroid c_n get closer

We want to minimize the negative log-likelihood, *i.e.* softmax loss ℓ:

$$\arg\min_{\theta} \ell = \arg\min_{\theta} -log(p), \tag{3}$$

where θ denotes the set of all parameters in CNN. Now we can infer the gradients of softmax loss $\ell_{\mathbf{f}}$ with respect to the anchor $\mathbf{w}_n$ given the single sample $\mathbf{f}$:

$$\nabla_{\mathbf{w}_n} \ell_{\mathbf{f}} = \frac{\partial \ell_{\mathbf{f}}}{\partial \mathbf{w}_n} = -\sum_{\mathbf{f}\in\mathcal{I}} \left(\mathbb{I}[\mathbf{f} \in \mathcal{I}_n] - \frac{\exp(\mathbf{y}_n)}{\sum_{i=1}^{N} \exp(\mathbf{y}_i)} \right) \cdot \mathbf{f}, \tag{4}$$

in which the samples of class n is denoted as $\mathcal{I}_n$, and $\mathbf{y}_n$ is the n^{th} element in $\mathbf{y}$. $\mathbb{I}$ refers to the indicator which is 1 when $\mathbf{f}$ is in $\mathcal{I}_n$, and 0 *vice versa.*

Now considering samples in one mini-batch, the gradient $\nabla_{\mathbf{w}_n} \ell$ with respect to results in the summation of all feature samples in the class n with a negative contribution from the summation of feature samples from the rest classes:

$$\nabla_{\mathbf{w}_n} \ell = -\sum_{\mathbf{f}\in\mathcal{I}_n} \left(1 - \frac{\exp(\mathbf{y}_n)}{\sum_{n=1}^{N} \exp(\mathbf{y}_n)} \right) \cdot \mathbf{f} + \sum_{\mathbf{f}\notin\mathcal{I}_n} \frac{\exp(\mathbf{y}_n)}{\sum_{n=1}^{N} \exp(\mathbf{y}_n)} \cdot \mathbf{f}.$$

In each iteration, the update value of $\mathbf{w}_n$ equal to

$$\Delta \mathbf{w}_n = -\eta \dot{\nabla}_{\mathbf{w}_n} \ell = \eta \sum_{\mathbf{f}\in\mathcal{I}_n} \left(1 - \frac{\exp(\mathbf{y}_n)}{\sum_{n=1}^{N} \exp(\mathbf{y}_n)} \right) \cdot \mathbf{f} - \eta \sum_{\mathbf{f}\notin\mathcal{I}_n} \frac{\exp(\mathbf{y}_n)}{\sum_{n=1}^{N} \exp(\mathbf{y}_n)} \cdot \mathbf{f}.$$

Where η denote the learning rate. The former term can be assumed as the scaled summation of the data samples in class n, thus is approximately proportional to the class centroid $\mathbf{c}_n$. And the feature samples are usually evenly distributed in the feature space, the summation of the negative feature samples for class n will also approximately follow the negative direction of the centroid $\mathbf{c}_n$. Therefore, the gradient $\nabla_{\mathbf{w}_n} \ell$ approximately points to the centroid direction $\mathbf{c}_n$ in one time step, thus finally the anchor $\mathbf{w}_n$ will also follow the direction of the centroid with sufficient accumulation of the gradients. Fig. 3 describes the moving direction of anchor $\mathbf{w}_n$ with the gradient $\Delta \mathbf{w}_n = -\nabla_{\mathbf{w}_n} \ell$ and the direction of samples x_n with the gradient $\Delta \mathbf{x}_n = -\nabla_{\mathbf{x}_n} \ell$ marked in red dot lines. For a class n, the

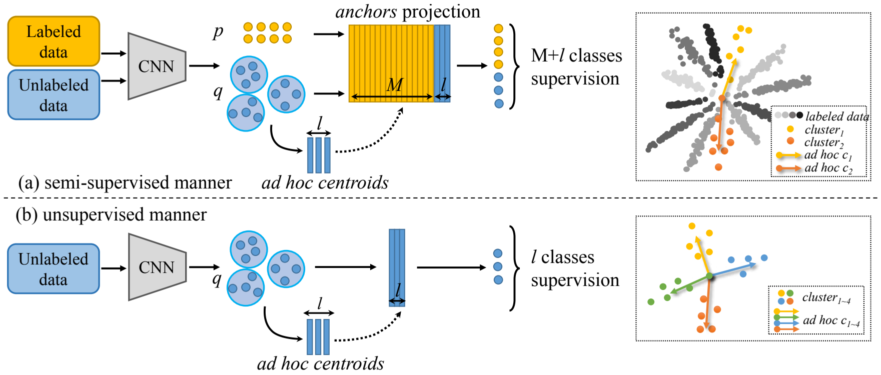


Fig. 4. A comparison between (a) semi-supervised learning with the proposed *transductive centroid projection* and (b) unsupervised learning framework

samples and anchors are marked with yellow dots and arrow line, respectively. When the network back-propagates, the direction of $\mathbf{w}_n$ is updated towards the class centroid $\mathbf{c}_n$ in tangential direction whilst the samples $\mathbf{x_n} \in \mathcal{I}_n$ are also gradually transformed to the direction of $\mathbf{w}_n$, which leads to $\sum_{j=1}^{o} x_{nj} = c_n \rightarrow w_n$.

3 Approach

Inspired by the observation stated in the previous section, we propose a novel learning mechanism to wisely congregate the unlabelled data into the recognition system to enhance its discriminative ability. Let $\mathcal{X}^{\mathrm{L}}$ denote the labelled dataset with M classes and $\mathcal{X}^{\mathrm{U}}$ the unlabelled dataset. We first cluster the $\mathcal{X}^{\mathrm{U}}$ by [24] and get N clusters. According to the property $w_n \approx c_n$ discussed in the previous section, the *ad hoc* centroid $\mathbf{c}^{\mathrm{U}}$ from an unlabelled cluster can be used to build up the corresponding *anchor* vector $\mathbf{w}^{\mathrm{U}}$, which means that it is possible to utilize the *ad hoc* centroid for a faithful classification of the unlabelled cluster.

3.1 Transductive Centroid Projection (TCP)

In one training step, we construct the mini-batch $\mathcal{B} = \{\mathcal{X}_p^{\mathrm{L}}, \mathcal{X}_q^{\mathrm{U}}\}$ by the labelled data $\mathcal{X}_p^{\mathrm{L}} \subset \mathcal{X}^{\mathrm{L}}$ and unlabelled data $\mathcal{X}_q^{\mathrm{U}} \subset \mathcal{X}^{\mathrm{U}}$, with $p = \mathrm{card}(\tilde{\mathcal{X}}^{\mathrm{L}})$ and $q = \mathrm{card}(\tilde{\mathcal{X}}^{\mathrm{U}})$ denote the number of selected labelled and unlabelled data in this batch, respectively. We randomly select $\mathcal{X}_p^{\mathrm{L}}$ from the labelled dataset as usual, but the unlabelled data are constructed by randomly selecting l unlabelled clusters with o samples in each cluster, *i.e.* $q = l \times o$. Note that the selected l clusters are dynamically changed for each mini-batch. Therefore, this mini-batch $\mathcal{B}$ is then fed into the network and the extracted features before the `TCP` layer are reformulated as $\mathbf{f} = [\mathbf{f}^{\mathrm{L}}, \mathbf{f}^{\mathrm{U}}]^{\top} \in \mathbb{R}^{(p+q)\times D}$, where D is the feature dimension and $\mathbf{f}^{\mathrm{L}}, \mathbf{f}^{\mathrm{U}}$ denote the feature vectors for labelled and unlabelled data, respectively.

The projection matrix for the `TCP` layer is reformulated as $\mathbf{W} = [\mathbf{W}^M, \mathbf{W}^l] \in \mathbb{R}^{(M+l)\times(p+q)}$, in which the first M columns are reserved for the anchors of labeled classes and the rest l columns are substituted by the *ad hoc* centroid vectors $\{\mathbf{c}^{\mathrm{U}}_\iota\}^l_{\iota=1}$ from the selected unlabeled data. Note that $\mathbf{c}^{\mathrm{U}}_\iota$ is calculated through the selected samples $\{\mathbf{f}^{\mathrm{U}}_{\iota,i}\}^o_{i=1}$ of the cluster ι in this mini-batch as

$$\mathbf{c}^{\mathrm{U}}_\iota = \alpha \sum_{i=1}^{o} \frac{\mathbf{f}^{\mathrm{U}}_{\iota,i}}{\|\mathbf{f}^{\mathrm{U}}_{\iota,i}\|_2}, \text{ where } \alpha = \frac{1}{M}\sum_{j=1}^{M} \|\mathbf{c}^{\mathrm{L}}_j\|_2. \tag{5}$$

The scale factor α is the average magnitude of the centroids for the labeled clusters. The output of the `TCP` layer is thereby obtained by $\mathbf{y} = \mathbf{W}^\top \mathbf{f}$ without the bias term, which is then fed into the `softmax` loss layer.

Compared to the training in a purely unsupervised manner, the semi-supervised learning procedure in this paper (as shown in Fig. 4(a)) applies the proposed transductive centroid projection layer which not only optimizes the inference towards the labeled data but also indirectly gains the recognition ability for the unlabeled clusters. Actually, it can be easily transferred to the unsupervised learning paradigm by setting $M = 0$ as shown in Fig. 4(b), or the supervised learning framework when there is no unlabeled data as $l = 0$.

3.2 Scale Factor α Matters

As stated in Sec. 3.1, the scale factor α is applied to normalize the *ad hoc* centroids for the unlabeled data. For the purpose of training stability and fast convergence, a suitable scaling criterion is to let the mapped activation $\mathbf{y}^{\mathrm{U}}$ for unlabeled data have a scale similar to the labeled one $\mathbf{y}^{\mathrm{L}}$. Indeed, the ℓ_2 norm of each centroid inherently offers a reasonable prior scale in mapping the input features $\mathbf{f}^{\mathrm{L}}$ to the output activation $\mathbf{y}^{\mathrm{L}}$. Therefore, by scaling the *ad hoc* centroids for the unlabeled data with an average scale $\alpha = \frac{1}{M}\sum_{j=1}^{M}\|\mathbf{c}^{\mathrm{L}}_j\|_2$ as the labeled centroids, activations for unlabeled data will have a similar distribution as the labeled activations, thus ensuring the stability and fast convergence during training.

3.3 Avoid Inter-class Conflict in Large Mini-batch

A larger batch size theoretically induces a better training performance in conventional recognition tasks. However, in TCP, it might be possible that a larger batch size will introduce multiple clusters with a same class label for the unlabelled data. Let the classes be evenly distributed in the unlabeled clusters, and assume that N clusters in the unlabelled data actually belong to $\tilde{N}$ classes, the probability that every cluster has a unique class label in the mini-batch $\mathcal{B}$ is $P(l) = (1 - \frac{N/\tilde{N}-1}{N})^l$, where l is the number of selected clusters. This probability decreases as the batch size increases, as shown in Fig. 5.

In our experiment, the ratio $N/\tilde{N} \simeq 8$ for person re-id and $N/\tilde{N} \simeq 3$ for face recognition. To guarantee the probability $P(l) > 0.99$, the number of cluster l

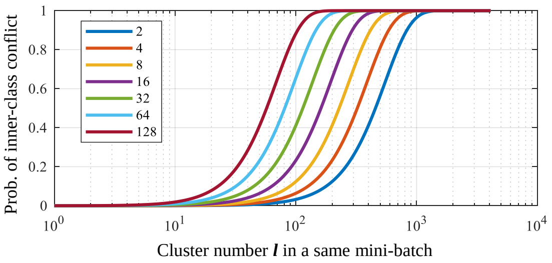


Fig. 5. The probability of each single cluster owning a unique class label in a mini-batch decreases with respect to the batch size. Seven ratios $N/\tilde{N}$ are marked in different colors

selected in a mini-batch should not be larger than 40. To further increase the number of unlabelled clusters in the mini-batch as much as possible, we provide two strategies as follows:

Selection of Clusters – Based on the assumption that the probability of inter-class conflict reduces along with the time interval during data collection, to avoid the conflict in training stage, the l clusters should be picked with an minimum interval T_l. In the experiment, we find that $T_l \geq 120$ seconds presents a good performance.

Selection of Samples – The diversity of samples extracted from consecutive frames in one cluster is always too small to aid intra-class feature learning. To this end, we make a constraint on sample selection by setting the interval between each sampled frame larger than T_o. In the experiment, we set T_o as 1 second.

Based on the aforementioned strategies, we find that only 19 out of $10,000$ mini-batches on Re-ID and 7 out of $10,000$ mini-batches on face recognition have duplicated identities when setting $l = 48$ in our training dataset.

3.4 Discussion: Stability and Efficiency

We further discuss the superiority of the proposed TCP layer comparing with some other metric learning losses, such as triplet loss [2] and contrastive loss [37], that can also avoid inter-class conflict by elaborate batch selection. Both of these losses suffer from dramatic data expansion when forming the sample pairs or sample triplets from the training set. Take triplet loss as an example, n unlabelled samples constitute $\frac{1}{3}n$ triplet sets and the metric only restricts on $\frac{2}{3}n$ distances in each iteration, *i.e.* the anchor to the negative sample and the anchor to the positive sample in each single triplet. It makes the triplet term suffer severe disturbance during training. Alternatively, in the proposed TCP layer, $n = p + q$ samples are compared with all the M anchors by labelled data as well as the l ad hoc centroids of the unlabeled data to achieve $(M + l) \times (p + q)$ comparisons, which is quadratically larger than other metric learning methods. It thus ensures a stable training process and a quick convergence.

Table 2. The list of eight datasets for training with their respective image and identity numbers

	CUHK03	CUHK01	PRID	VIPeR	3DPeS	i-LIDS	SenseReId	Market-1501	Total
# Tr. ID	1,467	971	385	632	193	119	16,377	751	20,895
# Tr. Imgs	21,012	1,552	2,997	506	420	194	160,396	10,348	197,425

4 Experimental Settings and Implementation details

Labeled Data and Unlabeled Data. For both of person re-identification and face recognition, the training data consist of two parts: labeled data $\mathcal{D}^{\mathrm{L}}$ and unlabeled data $\mathcal{D}^{\mathrm{U}}$.

In experiments for Re-ID, following the pipeline of DGD [38] and Spindle [39], we take the combined training samples from eight datasets described in Tab. 2 together as $\mathcal{D}^{\mathrm{L}}$. Note that MARS [13] is excluded from the training set since it is an extension of Market-1501. For $\mathcal{D}^{\mathrm{U}}$ construction, we collect videos with a total length of four hours from three different scenes with four cameras. The person clusters are obtained by the POI tracker [40] and clustered by [24] without further alignment, where those shorter than one second are removed. The unlabeled dataset, named as Person Tracker Re-Identification dataset (PT-ReID)[4], contains $158,446$ clusters and $1,324,019$ frames in total. For ablation study, we further manually annotate the PT-ReID, named as Labeled PT-ReID dataset (L-PT-ReID), and get a total of $2,495$ identities.

In experiments for face recognition, we combine a labelled MS-Celeb-1M [35] with some collected photos from internet as $\mathcal{D}^{\mathrm{L}}$, which in total contains $\sim 10M$ images and $1.6M$ identities. For $\mathcal{D}^{\mathrm{U}}$ we collect $11.0M$ face frames from surveillance videos and cluster them into $500K$ clusters. All faces are detected and aligned by [41].

Evaluation Benchmarks. For Re-ID, The proposed method is evaluated on six significant publicly benchmarks, including the image-based Market-1501 [42], CUHK01 [43], CUHK03 [44], and the video-based MARS [13], iLIDS-VID [45] as well as Prid2011 [46]. For face recognition, we evaluate the method on NIST IJB-C [47], which contains 138000 face images, 11000 face videos, and 10000 non-face images. To the best of our knowledge, it is the latest and the most challenging benchmarks for face verification. Notice that we found more than one hundred wrong annotations in this dataset, which introduce significant confusion for recall rate on some small false positive rate (FPR $\leq$ 1e-3), so we remove these pairs in evaluation[5].

Evaluation Metrics. For Re-ID, the widely used Cumulative Match Curve (CMC) is adopted in both ablation study and comparison experiments. In addition, we apply Mean Average Precision (MAP) as another metric for evaluations on Market-1501 [42] and MARS [13] dataset. For face recognition, the receiver operating characteristic (ROC) curve is adopted as in most of the other works.

[4] The dataset will be released.

[5] The list will be made available.

Table 3. Comparison results of different baselines with the proposed TCP (last row) on Market-1501 dataset. All pipelines are trained by a plain ResNet-101 without any bells and whistles. The top four are single-task learning with single data source (*i.e.* $\mathcal{D}^{\mathrm{L}}$ or $\mathcal{D}^{\mathrm{U}}$), while the following five take both data sources with multi-task learning

Methods	**Top-1**	**Top-5**	**Top-10**	**Top-20**	**MAP**
$\mathbf{S}^{\mathrm{L}}$	87.7	93.5	95.1	96.6	79.4
$\mathbf{S}^{\mathrm{U}}$	22.8	32.2	36.6	41.8	8.6
$\mathbf{S}^{\mathrm{U}}_{\mathrm{self}}$	65.0	77.0	82.9	93.5	61.3
$\mathbf{S}^{\mathrm{U}}_{\mathrm{labeled}}$	66.4	78.0	83.4	98.0	67.6
$\mathbf{M}^{\mathrm{U+L}}$	37.4	46.6	51.5	67.0	21.0
$\mathbf{M}^{\mathrm{U+L}}_{\mathrm{self}}$	68.8	79.9	84.6	94.5	55.0
$\mathbf{M}^{\mathrm{U+L}}_{\mathrm{labeled}}$	86.0	90.8	92.7	94.8	75.8
$\mathbf{M}^{\mathrm{U+L}}_{\mathrm{tr\text{-}loss}}$	83.5	89.5	93.5	95.9	79.3
$\mathbf{M}^{\mathrm{U+L}}_{\mathrm{TCP}}$	89.6	94.1	95.6	96.8	83.5
TCP	**90.4**	**94.5**	**95.7**	**96.9**	**84.4**

On all datasets, we compute cosine distance between each pair of query image and any image from the gallery, and return the ranked gallery list.

Training Details. As a common practice in most deep learning frameworks for visual tasks, we initialize our model with the parameters pre-trained on ImageNet. Specifically, we employ resnet-101 as the backbone structure in all experiments which is followed by an additional `fc` layer after `pool5` to generate 128-D features. Dropout [48] is used to randomly drop out a channel with the ratio of 0.5. The input size is normalized as 224×224 and the training batch size is $3,840$, in which $p = 2,880, q = 960, l = 96$ and $o = 10$. Warm up technology [49] is used to achieve stability when training with large batch size.

5 Ablation Study

Since the training data, network structure and pre-processing for the data vary from method to method, we first analyse the effectiveness of the proposed method with quantitative comparisons to different baselines in Sec. 5.1 and visualize the feature space in Sec. 5.2. All the ablation study are conducted on Market-1501, a large-scale clean dataset with strong generalizability.

5.1 Component Analysis

Since the semi-supervised learning contains two data sources, *i.e.* labeled data $\mathcal{D}^{\mathrm{L}}$ and unlabeled data $\mathcal{D}^{\mathrm{U}}$, the proposed TCP is compared with nine typical configuration baselines listed in Tab. 3. These baselines can be divided into two types: single-task learning with only one data source and multi-task learning with multiple data sources.

The top four are single-task learning with single data source: (1) $\mathbf{S}^{\mathrm{L}}$ only uses $\mathcal{D}^{\mathrm{L}}$ supervised by the annotated ground truth IDs with softmax loss; (2) $\mathbf{S}^{\mathrm{U}}$ only uses $\mathcal{D}^{\mathrm{U}}$ supervised by taking the cluster IDs as the pseudo ground truth with softmax loss; (3) $\mathbf{S}^{\mathrm{U}}_{\mathrm{self}}$ with self-training on unlabeled data, where self-training is a classical semi-supervised learning method. We first train the CNN with $\mathcal{D}^{\mathrm{L}}$ which is used to extract features of $\mathcal{D}^{\mathrm{U}}$, and then obtain the pseudo ground truth by a cluster algorithm. The pseudo ground truth is taken as the supervision for training on $\mathcal{D}^{\mathrm{U}}$; and (4) $\mathbf{S}^{\mathrm{U}}_{\mathrm{labeled}}$ - We further annotate the real ground truth of unlabeled data and compare it with the model trained with pseudo ground truth.

The latter five are multi-task learning and three of them are a combination of the above single-task baselines as follows: (5) $\mathbf{M}^{\mathrm{U+L}}$ combines $\mathbf{S}^{\mathrm{L}}$ and $\mathbf{S}^{\mathrm{U}}$; (6) $\mathbf{M}^{\mathrm{U+L}}_{\mathrm{self}}$ is a combination of $\mathbf{S}^{\mathrm{L}}$ and $\mathbf{S}^{\mathrm{U}}_{\mathrm{self}}$; and (7) $\mathbf{M}^{\mathrm{U+L}}_{\mathrm{labeled}}$ is a combination of $\mathbf{S}^{\mathrm{L}}$ and $\mathbf{S}^{\mathrm{U}}_{\mathrm{labeled}}$. The last two take the annotated ground truth to supervise the branch with labeled data and compare the performance of operating triplet loss with our TCP on unlabeled data as (8) $\mathbf{M}^{\mathrm{U+L}}_{\text{tr-loss}}$ with triplet loss, where the selection strategy for triplets also follow the Online Batch Selection described in Sec.3.3, and (9) $\mathbf{M}^{\mathrm{U+L}}_{\mathrm{TCP}}$ utilizes the proposed TCP which is regarded as training in a unsupervised manner.

The proposed method TCP is neither single-task nor multi-task learning, but with the labeled and unlabeled data trained simultaneously in a semi-supervised manner. The results clearly prove that either single-task or multi-task learning will pull down the performance which are concluded as follows:

Clustered data contain noisy and fake ground truth. Compared with the näive baseline $\mathbf{S}^{\mathrm{U}}$ that directly uses cluster IDs as the supervision, the self-training $\mathbf{S}^{\mathrm{U}}_{\mathrm{self}}$ outperforms it by 42%. Similarly, by fusing labeled data, the $\mathbf{M}^{\mathrm{U+L}}_{\mathrm{self}}$ is superior to $\mathbf{M}^{\mathrm{U+L}}$ with 31.4%. It shows that (1) the source cluster data contains many fake ground truth and (2) many cluster fragments cause the same identity to be clustered to different ID ground truth.

It's hard to manually refine unlabelled cluster data. We further annotate the cluster data to get the real ground truth of unlabeled data. Although $\mathbf{S}^{\mathrm{U}}_{\mathrm{labeled}}$ outperforms $\mathbf{S}^{\mathrm{U}}$ with pseudo ground truth again demonstrating the noise of cluster, both $\mathbf{S}^{\mathrm{U}}_{\mathrm{labeled}}$ and $\mathbf{M}^{\mathrm{U+L}}_{\mathrm{labeled}}$ drop performance compared to training on labeled data $\mathbf{S}^{\mathrm{L}}$. It shows that there is a significant disparity between two source data domains, and it is non-trivial to get a clean annotation set due to the time gap between different clusters.

Self-training and triplet-loss are not optimal. Both self-training $\mathbf{M}^{\mathrm{U+L}}_{\mathrm{self}}$ and triplet-loss $\mathbf{M}^{\mathrm{U+L}}_{\text{tr-loss}}$ provide solutions to overcome the problems caused by the pseudo ground truth of clusters data, significantly performing the näive combination of unlabeled and labeled data $\mathbf{M}^{\mathrm{U+L}}$, however, their results are still lower than that of our method by 21.6% and 6.9% respectively. As discussed in Sec. 3.4, the triplet-loss only consider $\frac{2}{3}N$ distances that cannot fully exploit the information in each batch data, while self-training profoundly depends on the robustness of the pre-trained model with labeled data that cannot be guaranteed to intrinsically solve the problem.

The superiority of TCP. By employing TCP, both the unsupervised learning $\mathbf{M}^{\mathrm{U+L}}_{\mathrm{TCP}}$ and semi-supervised learning TCP, not surprisingly, outperform all of the above baseline variants by a large margin. It proves the superiority of the proposed online batch selection and the centroid projection mechanism which

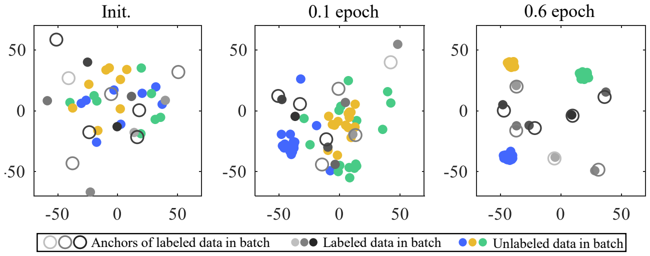


Fig. 6. Feature and anchor distribution converge during semi-supervised training with the proposed TCP layer

comprehensively utilize all labeled as well as unlabeled data by optimizing $(M + l) \times (p + q)$ distances.

5.2 Feature Hyperspace on Person Re-ID

The feature spaces learned on MNIST, CIFAR-100 and MS1M are discussed in Sec. 2.1. Here we examine whether the same observations and conclusions also occur on person re-identification with the proposed TCP layer, by visualizing the distribution related to the mini-batches on a single GPU in different training stages. For a clear visualization, we show the mini-batch with 8 labeled samples where each belongs to a distinct class and 24 unlabeled samples from 3 classes each of which has 8 samples in Fig. 6. As the number of epoch increases, the anchors of labeled data converge towards their corresponding sample centroids while those of unlabeled data keep still in the centroids. Until the network converges, the anchors of both labeled and unlabeled data are in the centroid of each class and thus the unlabeled data can be regarded as the auto-annotated data to enlarge the training data span.

6 Evaluation on Seven Benchmarks

6.1 Person Re-identification Benchmarks

We first evaluate our method on the six Re-ID benchmarks. Notice that since the data pre-processing, training setting and network structure vary in different state-of-the-art methods, we only list recent best performing methods in the tables just for reference. The test procedure on iLIDS-VID and PRID2011 is the average of 10-fold cross validation result, whereas on MARS we use a fixed split following the official protocol [13]. As shown in Tab. 4, 'Basel.' denote the $\mathbf{S}^{\mathrm{L}}$ setting in 5. The proposed TCP, compared with a variety of recent methods, achieves the best performance on the Market-1501, CUHK03 and CUHK01 datasets. The performance will be further improved with an additional re-rank skill.

Table 4. Experimental results (%) of the proposed and other comparisons on six person re-identification datasets. The best are in bold while the second best are underlined

Market1501	**Top-1**	**Top-5**	**Top-10**	**Top-20**	**MAP**	**CUHK01**	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Best [50]	84.1	92.7	94.9	96.8	63.4	Best [39]	79.9	94.4	97.1	98.6
Basel.	82.7	92.3	95.0	96.0	58.1	Basel.	83.0	96.2	98.1	99.3
TCP	86.1	94.0	95.0	96.2	66.2	TCP	90.0	98.0	99.0	**99.4**
TCP + Re-rank	**90.4**	**94.5**	**95.7**	**96.9**	**84.4**	TCP + Re-rank	**91.6**	**98.3**	**99.1**	**99.4**
MARS	**Top-1**	**Top-5**	**Top-10**	**Top-20**	**MAP**	**iLIDS-VID**	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Best [51]	73.9	-	-	-	**68.4**	Best [52]	62.0	86.0	94.0	98.0
Basel.	77.2	90.4	93.3	95.1	47.7	Basel.	64.5	91.8	96.9	98.8
TCP	80.7	91.6	**94.4**	95.7	53.7	TCP	69.4	**95.1**	**98.3**	**99.3**
TCP + Re-rank	**82.9**	**91.8**	93.7	96.4	67.6	TCP + Re-rank	**71.7**	**95.1**	**98.3**	**99.3**
CUHK03	**Top-1**	**Top-5**	**Top-10**	**Top-20**	-	**PRID2011**	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Best [50]	88.7	98.6	99.2	99.6	-	Best [52]	77.0	95.0	99.0	99.0
Basel.	91.7	99.1	99.6	99.8		Basel.	84.6	95.4	99.0	99.6
TCP	94.4	99.7	99.9	**100.0**	-	TCP	92.1	98.1	**99.6**	**100.0**
TCP + re-rank	**98.2**	**100.0**	**100.0**	**100.0**	-	TCP + Re-rank	**93.6**	**98.9**	**99.6**	**100.0**

Table 5. Experimental results (%) on IJB-C and LFW datasets

Benchmark	**IJB-C**							**LFW**
Index	tpr@1e-1	tpr@1e-2	tpr@1e-3	tpr@1e-4	tpr@1e-5	tpr@1e-6	tpr@1e-7	Acc
Best [32]	-	-	-	-	-	-	-	99.80
$\mathbf{S}^{\mathrm{U}}$	98.65	95.08	84.14	64.98	40.42	21.89	9.94	98.24
$\mathbf{S}^{\mathrm{L}}$	99.70	98.98	97.37	94.62	90.49	83.68	76.37	99.78
$\mathbf{S}^{\mathrm{U+L}}_{\mathrm{self}}$	98.97	98.80	98.16	96.60	93.67	88.64	80.69	99.80
TCP	**99.97**	**99.81**	**99.16**	**97.58**	**94.63**	**89.21**	**82.90**	**99.82**

6.2 Face Recognition Benchmarks

IJB-C [47] is the most challenging face recognition benchmark for now. Since it has just been released for a few months, few work report its result on it. We report the true positive rates on seven different levels of false positive rates (from 1e-1 to 1e-7) in Fig. 5. Comparison has been made between the proposed TCP with some baselines as discribed in Sec. 5. The best accuracy of existing works on the widely used LFW dataset is also reported for reference. The result of the proposed TCP outperforms all the baselines especially the self-training one, the training process of which takes more than 4-times the time of TCP.

7 Conclusion

By observing the latent space learned by softmax loss in CNN, we propose a semi-supervised method named TCP which can be steadily embedded in CNN and followed by any classification loss functions. Extensive experiments and ablation study demonstrate its superiority in utilizing full information across labelled and unlabelled data to achieve state-of-the-art performance on six person re-identification datasets and one face recognition dataset.

References

1. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 1701–1708
2. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 815–823
3. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1891–1898
4. Sun, Y., Liang, D., Wang, X., Tang, X.: Deepid3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
5. Liu, Y., Li, H., Wang, X.: Rethinking feature discrimination and polymerization for large-scale recognition. arXiv preprint arXiv:1710.00870 (2017)
6. Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S.: Region-based quality estimation network for large-scale person re-identification. arXiv preprint arXiv:1711.08766 (2017)
7. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: CVPR. Volume 2. (2017) 8
8. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
9. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T.: See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
10. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
11. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. arXiv preprint arXiv:1705.04724 (2017)
12. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In: European Conference on Computer Vision, Springer (2016)
13. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: European Conference on Computer Vision, Springer (2016) 868–884
14. Weston, J., Ratle, F., Mobahi, H., Collobert, R.: Deep learning via semi-supervised embedding. In: Neural Networks: Tricks of the Trade. Springer (2012) 639–655
15. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML. Volume 3. (2013) 2
16. Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., Bu, J.: Semi-supervised coupled dictionary learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 3550–3557
17. Odena, A.: Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583 (2016)
18. Fan, H., Zheng, L., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. arXiv preprint arXiv:1705.10444 (2017)

19. Yang, J., Parikh, D., Batra, D.: Joint unsupervised learning of deep representations and image clusters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 5147–5156
20. Wang, X., Lu, L., Shin, H.C., Kim, L., Bagheri, M., Nogues, I., Yao, J., Summers, R.M.: Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition. In: Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, IEEE (2017) 998–1007
21. Zhu: Learning from labeled and unlabeled data with label propagation. (2002)
22. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Volume 1., Oakland, CA, USA (1967) 281–297
23. Gowda, K.C., Krishna, G.: Agglomerative clustering using the concept of mutual nearest neighbourhood. Pattern recognition **10**(2) (1978) 105–112
24. Gdalyahu, Y., Weinshall, D., Werman, M.: Self-organization in vision: stochastic clustering for image segmentation, perceptual grouping, and image database organization. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(10) (2001) 1053–1074
25. Kurita, T.: An efficient agglomerative clustering algorithm using a heap. Pattern Recognition **24**(3) (1991) 205–209
26. Cozman, F.G.: Semi-supervised learning of mixture models. In: ICML. (2003)
27. Bennett, K.P.: Semi-supervised support vector machines. In: NIPS. (1999) 368–374
28. Liu, W., Wang, J., Chang, S.F.: Robust and scalable graph-based semisupervised learning. Proceedings of the IEEE **100**(9) (2012) 2624–2638
29. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
30. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
32. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. Volume 4. (2017) 12
33. Lecun, Y., Cortes, C.: The mnist database of handwritten digits. (2010)
34. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. (2009)
35. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. Electronic Imaging **2016**(11) (2016) 1–6
36. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(Nov) (2008) 2579–2605
37. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Computer vision and pattern recognition, 2006 IEEE computer society conference on. Volume 2., IEEE (2006) 1735–1742
38. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1249–1258
39. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindlenet: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1077–1085

40. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: Poi: Multiple object tracking with high performance detection and appearance feature. In: European Conference on Computer Vision, Springer (2016) 36–42
41. Liu, Y., Li, H., Yan, J., Wei, F., Wang, X., Tang, X.: Recurrent scale approximation for object detection in cnn. In: IEEE international conference on computer vision. Volume 5. (2017)
42. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1116–1124
43. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: ACCV. (2012)
44. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR. (2014)
45. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: European Conference on Computer Vision, Springer (2014) 688–703
46. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person Re-Identification by Descriptive and Discriminative Classification. In: Proc. Scandinavian Conference on Image Analysis (SCIA). (2011)
47. : The iarpa janus benchmark-c face challenge (ijb-c). https://www.nist.gov/programs-projects/face-challenges Accessed: 2018-03-315.
48. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of machine learning research **15**(1) (2014) 1929–1958
49. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
50. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
51. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
52. Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)