Catalyst for Clustering-Based Unsupervised Object Re-identification: Feature Calibration

Huafeng Li[1*], **Qingsong Hu**[1*], **Zhanxuan Hu**[2†]

[1]School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China

[2]School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China

hfchina99@163.com, qingsonghu08@gmail.com, zhanxuanhu@gmail.com

Abstract

Clustering-based methods are emerging as a ubiquitous technology in unsupervised object Re-Identification (ReID), which alternate between pseudo-label generation and representation learning. Recent advances in this field mainly fall into two groups: pseudo-label correction and robust representation learning. Differently, in this work, we improve unsupervised object ReID from feature calibration, a completely different but complementary insight from the current approaches. Specifically, we propose to insert a conceptually simple yet empirically powerful Feature Calibration Module (FCM) before pseudo-label generation. In practice, FCM calibrates the features using a nonparametric graph attention network, enforcing similar instances to move together in the feature space while allowing dissimilar instances to separate. As a result, we can generate more reliable pseudo-labels using the calibrated features and further improve subsequent representation learning. FCM is simple, effective, parameter-free, training-free, plug-and-play, and can be considered as a catalyst, increasing the 'chemical reaction' between pseudo-label generation and representation learning. Moreover, it maintains the efficiency of testing time with negligible impact on training time. In this paper, we insert FCM into a simple baseline. Experiments across different scenarios and benchmarks show that FCM consistently improves the baseline (e.g., 8.2% mAP gain on MSMT17), and achieves the new state-of-the-art results. Code is available at: https://github.com/lhf12278/FCM-ReID.

Introduction

Object Re-identification (ReID) aims to retrieve the same object across different time stamps, locations, and cameras (Ye et al. 2021). Although supervised object ReID has achieved tremendous success in multiple different scenarios (Luo et al. 2019; Jiang and Ye 2023), the requirement of enormous labels limits its practicality in the real world. An alternative solution to alleviate the need for labeled data is unsupervised object ReID which trains the feature extraction network using only unlabeled data. Particularly, clustering-based methods are emerging as a ubiquitous technology in recent years.

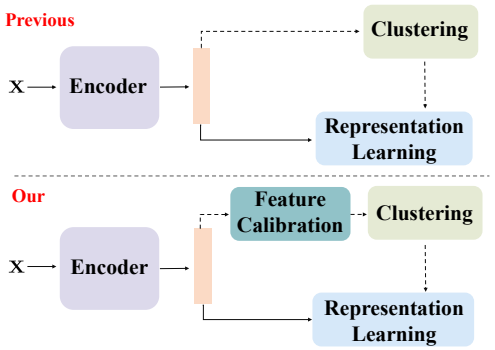


Figure 1: Conceptual illustration of our proposed learning paradigm for unsupervised object ReID. We improve the previous methods by introducing a *nonparametric feature calibration* module.

As shown in Figure1 (top), a general pipeline for previous clustering-based ReID methods is to alternate between clustering-based pseudo-label generation and pseudo-label-based representation learning. Essentially, these two steps are interdependent. On the one hand, pseudo-labels provide weakly supervised information that guides the representation learning model to extract discriminative features. On the other hand, discriminative features enable the clustering to generate more reliable pseudo labels. Recent studies improve unsupervised ReID mainly along the following two lines: pseudo-label correction and robust representation learning. The former aims to improve the pseudo-label quality using additional information, such as neighbor information (Yan et al. 2022) or part information (Cho et al. 2022). While the latter focuses on designing robust representation learning algorithms to mitigate the effect of noisy labels, such as LP (Lan et al. 2023) and ISE (Zhang et al. 2022c).

Differently, in this work, we enhance unsupervised ReID through feature calibration, a completely different but complementary insight from the current methods. As shown in Figure 1 (bottom), we propose to insert a Feature Calibration Module (FCM) before clustering-based pseudo-label generation. The main goal of FCM is to refine the features extracted from the model and further improve the clustering quality. In practice, we realize FCM by explor-

[*]These authors contributed equally.

[†]Corresponding author

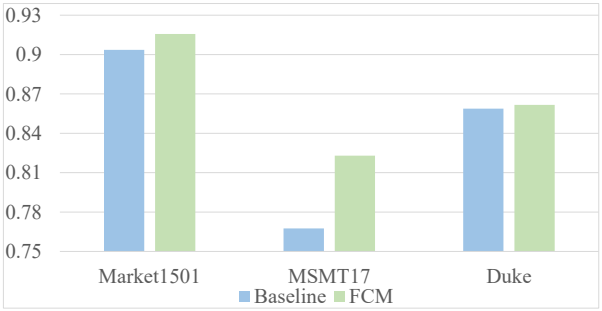


Figure 2: Clustering quality (AMI (Pedregosa et al. 2011)) comparison between the original features and calibrated features on Market-1501, MSMT17, and DukeMTMC-reID.

ing the sample relationships. More specifically, FCM builds upon a non-parametric graph attention network that constructs smooth, clustering-friendly node representations via aggregating neighborhood information. Notably, FCM is parameter-free, training-free, and can be abandoned in the testing stage, such that it enjoys both high training efficiency and high testing efficiency. Furthermore, FCM is plug-and-play and can be integrated into existing methods. In summary, the main contributions of this work can be summarized as follows:

1. *New perspective.* To the best of our knowledge, this work is the first study to explore and demonstrate the efficacy of feature calibration for clustering-based unsupervised object ReID. As shown in Figure 2, feature calibration consistently improves the clustering performance across different datasets.
2. *Novel approach.* We develop a conceptually simple yet empirically powerful feature calibration module, which can be easily inserted into the current clustering-based unsupervised object ReID methods, serving as a catalyst to increase the 'chemical reaction' between pseudo-label generation and representation learning.
3. *State-of-the-art results.* We perform extensive experiments to validate the effectiveness of our proposed feature calibration module. Experimental results show that it can significantly improve the performance of baselines and achieve a new state-of-the-art. For example, our method achieves 49.1% mAP on MSMT17 outperforms the best competitor by 6.0%.

Related Work

Unsupervised ReID

Clustering-based methods are emerging as a ubiquitous technology for unsupervised object ReID (Zeng et al. 2020a; Zheng et al. 2021; Jin et al. 2022; Dai et al. 2022a) and unsupervised domain adaptive ReID (Ge et al. 2020; Li et al. 2023). State-of-the-art methods generally involve two steps: clustering-based pseudo-label generation and pseudo-label-based representation learning. And, recent advances in this field can be roughly grouped into (i) pseudo-label correction, and (ii) robust representation learning.

The goal of pseudo-label correction is to enhance the accuracy of pseudo-labels generated through conventional clustering algorithms, such as DBSCAN (Ester et al. 1996) and K-means (Lloyd 1982). For instance, RPL (Zhang et al. 2021) employs a temporal ensembling approach to refine these pseudo-labels. PPLR (Cho et al. 2022) improves pseudo-labels by examining the complementary relationship between global and part features. Meanwhile, GLC (Yan et al. 2022) introduces a graph-based pseudo-label correction network that investigates the relationships between samples and their k-nearest neighbors. A similar idea has been studied in (Cheng et al. 2022).

Robust representation learning focus on designing an effective representation learning algorithm for extracting discriminative features, even when confronted with noisy pseudo-labels. MMCL (Wang and Zhang 2020) employs a memory-based non-parametric classifier, conducting both multi-label and single-label classification simultaneously. CC (Dai et al. 2022a) presents a simple cluster contrast-based method that maintains a cluster-level memory dictionary and conducts contrastive learning at the cluster level. Subsequently, CACL (Li, Li, and Guo 2022) improves this approach by simultaneously applying instance-level and cluster-level contrastive learning. To alleviate the effect of noisy clusters, ISE (Zhang et al. 2022c) generates support samples in the embedding space and introduces an additional label-preserving loss. LP (Lan et al. 2023) introduces a teacher-student network to reduce the interference of noisy labels during training. DCDP (Chen et al. 2023) generates two sets of pseudo labels using two separate networks and introduces a consistent sample mining strategy. Besides, HNGN (Li et al. 2022) incorporates a ReID network and a hard negative generation network into a joint framework, training them alternately using an adversarial approach.

Representation Learning with Sample Relationship Exploring

Several recent studies have concentrated on enhancing representation learning by investigating sample relationships. For instance, Batchformer (Hou, Yu, and Tao 2022) introduces a simple transformer into the batch dimension to capture and model the sample relationships within each mini-batch. NFormer (Wang et al. 2022) utilizes a landmark agent attention module to model the relation map between images, along with a reciprocal neighbor softmax mechanism to attain sparse attention, thereby alleviating the computational overhead. LIBC (Seidenschwarz, Elezi, and Leal-Taixé 2021) leverages a GNN-based message-passing network that considers the overall structure of a mini-batch. Besides, for the exposure correction task, ERL (Huang et al. 2023) proposes to correlate and constrain the sample relationship of the correction procedure in a minibatch. In the image-text matching task, HREM (Fu et al. 2023) explores semantic relationships among instances. However, all the aforementioned methods entail training an elaborate graph or transformer subnetwork, which introduces additional network parameters and hyperparameters. In contrast, our proposed FCM is parameter-free, training-free, and can be seamlessly integrated into existing clustering-based un-

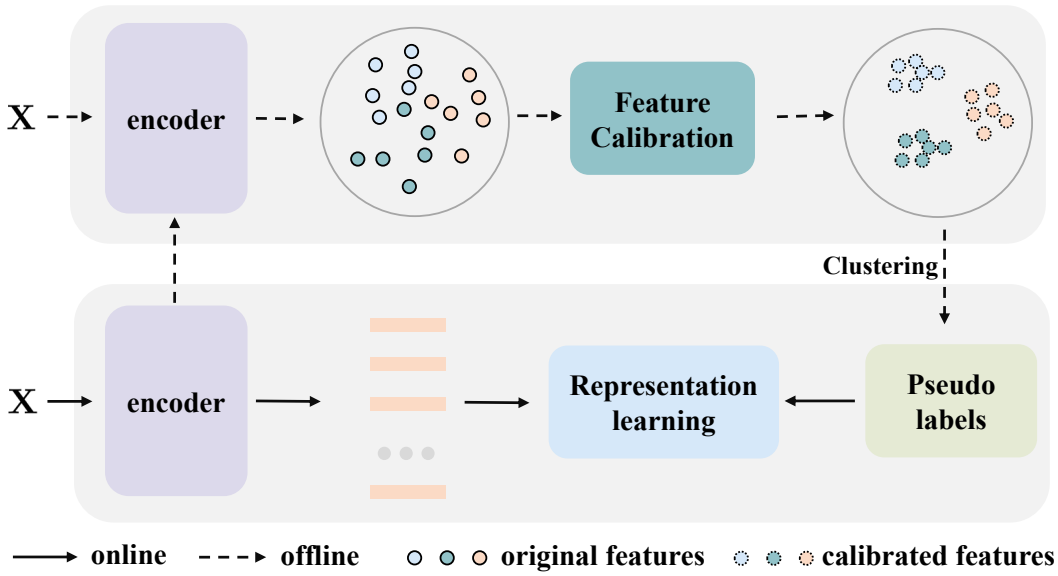


Figure 3: Pipeline of improved clustering-based unsupervised object ReID with feature calibration. We propose to insert a simple, parameter-free, and training-free feature calibration module before clustering-based pseudo-label generation.

supervised object ReID methods.

Method

In this section, we first provide an overview of the improved clustering-based unsupervised object ReID framework with feature calibration. Then, we introduce the key modules, including feature calibration, pseudo-label generation, and representation learning in detail.

Overview

We provide a general pipeline for improved clustering-based unsupervised object ReID with feature calibration in Figure 3. Suppose $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is the training data. At the start of each epoch, we sequentially perform the following four steps:1) *feature extraction*, we extract the features using an encoder $f_\theta(x)$, i.e., $\mathbf{f}_i = f_\theta(x_i)$; 2) *feature calibration*, we conduct feature calibration $\mathcal{C}$ to refine the features, i.e., $\mathbf{c}_i = \mathcal{C}(\mathbf{f}_i)$; 3) *pseudo-label generation*, we perform clustering on calibrated features $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n\}$ to generate the pseudo-labels $\{y_1, y_2, \ldots, y_n\}$; 4) *Representation learning*, we conduct representation learning with pseudo-labels to train the encoder $f_\theta(x)$. In practice, any backbone can be used in *feature extraction*. Next, we elaborate on the residual three steps.

Feature Calibration

Existing clustering-based object ReID methods often directly use the output of the encoder $f_\theta(x)$ to generate pseudo-labels. However, the quality of features is generally poor, especially in the early stages of training, as shown in Figure 4a. As a result, clustering will generate numerous incorrect pseudo-labels (as shown in Figure 2), misleading the training of encoder $f_\theta(x)$. And, this forms a vicious circle. To mitigate this issue, we propose to calibrate the features before clustering. Specifically, we achieve feature calibration using a nonparametric graph attention network. For clarity, we first introduce the traditional parametric graph attention networks (Velickovic et al. 2017).

Graph Attention Networks (GAT) is stacked by a series of Self-Attention (SA) layers in which node features are updated by aggregating their neighborhood's information. Given a set of node features $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$, a SA layer first computes the attention coefficients by:

$$e_{ij} = \mathcal{A}(\mathbf{W}\mathbf{f}_i, \mathbf{W}\mathbf{f}_j)\,, \tag{1}$$

where $\mathbf{W}$ denotes a shared linear transformation and works on all nodes, $\mathcal{A}$ denotes a shared attention mechanism. e_{ij} indicates the importance of $\mathbf{f}_j$ to $\mathbf{f}_i$. In practice, to achieve coefficients comparison across different nodes, $\{e_{ij}\}_{i,j=1:n}$ are normalized using the Softmax function, i.e.,

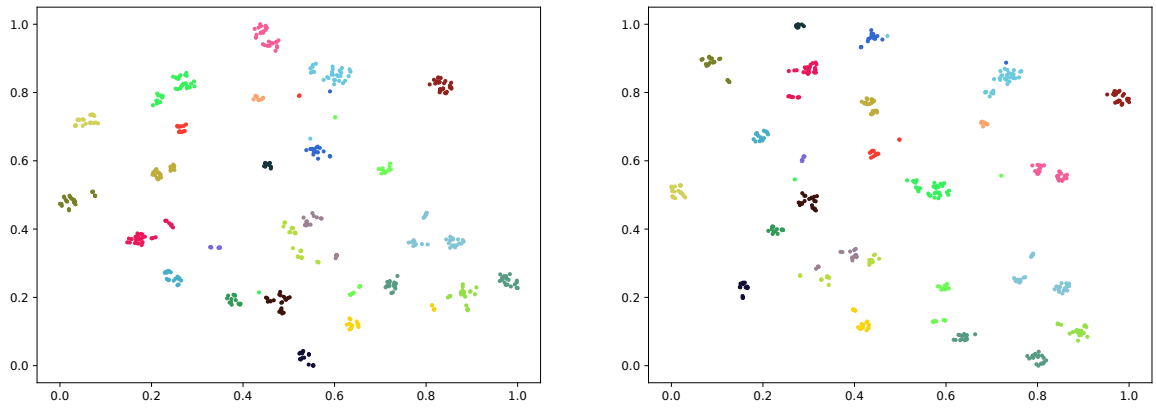
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j \in \mathcal{N}_i} \exp(e_{ij})}\,, \tag{2}$$

where $\mathcal{N}_i$ denotes the neighbors of i-th node. In practice, $\mathcal{A}$ is realized by a single-layer feedforward neural network and parametrized by a weight vector $\hat{\mathbf{a}}$. Besides, a LeakyReLU nonlinear layer is applied. Then, the normalized attention coefficients are expressed as

$$\alpha_{ij} = \frac{exp(\text{LeakyReLU}(\hat{\mathbf{a}}^T[\mathbf{W}\mathbf{f}_i \| \mathbf{W}\mathbf{f}_j]))}{\sum_{j \in \mathcal{N}_i} \exp(\text{LeakyReLU}(\hat{\mathbf{a}}^T[\mathbf{W}\mathbf{f}_i \| \mathbf{W}\mathbf{f}_j]))}\,, \tag{3}$$

where $\|$ denotes the concatenation operation. For i-th node, the output node feature is a linear combination of its neighbor's features using the normalized attention coefficients,

$$\vec{\mathbf{f}}_i = \sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\mathbf{f}_j\,. \tag{4}$$



(a) Original features (b) Calibrated features

Figure 4: t-SNE visualization of (a) features extracted by the model pre-trained on LUPerson and (b) calibrated features by our proposed FCM. Obviously, the calibrated features have a relatively clear clustering structure.

Intuitively, each node in GAT iteratively smooths the features of its neighbors for better node embedding. Compared with GCN, attention architecture enjoys two interesting properties: 1) it is computationally efficient due to involving only simple operations; 2) it is robust to the graph structure due to taking only a set of node features as input. *However, the training parameters prevent GAT from our proposed offline feature calibration.* To this end, we propose a nonparametric graph attention network.

Nonparametric Graph Attention Networks (NpGAT). As mentioned above, each SA layer contains two operations: node feature transformation and node feature aggregation. Recently, several studies (Zhang et al. 2022b,a) propose to decouple the feature aggregation and feature transformation in each GNN layer for scalable node classification. Concretely, they first execute the feature aggregation operation and then feed the aggregated features into a simple MLP to produce the final labels. Experiments demonstrate that these methods can achieve comparable or even better results than the one of coupled GNNs. Motivated by this observation, we hypothesize that the majority of the benefit in GAT arises from the feature aggregation, and the linear transformation in Eq (1) is not critical. Hence, we remove the linear transformation functions in each SA layer and only use a simple nonparametric attention mechanism. The resulting model is parameter-free and training-free, yet it retains the same information aggregation capabilities as GAT.

Specifically, given a set of features $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$ extracted from the model, we regard each sample as a node, and calculate the attention coefficients by

$$\hat{e}_{ij} = \cos(\mathbf{f}_i, \mathbf{f}_j)\,, \tag{5}$$

Similarly, we normalize it using the Softmax function, i.e.,

$$\hat{\alpha}_{ij} = \frac{\exp(\hat{e}_{ij}/\sigma)}{\sum_{j\in\mathcal{N}_i}\exp(\hat{e}_{ij}/\sigma)}\,, \tag{6}$$

where σ is a temperature parameter. Then, we aggregate the features by

$$\mathbf{c}_i = \sum_{j=1,\ldots,n} \hat{\alpha}_{ij}\mathbf{f}_j\,. \tag{7}$$

Notably, we can build a deep NpGAT via stacking multiple nonparametric SA layers. However, experiments verify that one SA layer is enough for generating sufficiently good pseudo labels.

Clustering-Based Pseudo-Label Generation

Two widely used clustering algorithms for pseudo-label generation are DBSCAN (Ester et al. 1996) and K-means (Lloyd 1982). In this work, we follow Cluster-Contrast (CC) (Dai et al. 2022a) and perform DBSCAN to group the calibrated features and generate pseudo-labels $\{y_1, y_2, \ldots, y_n\}$. Two hyper-parameters in DBSCAN are ϵ, the neighborhood radius, and N_r, the minimum number of samples in the neighborhood around a sample point. We set them to be consistent with CC for a fair comparison.

Representation Learning

At the representation learning stage, we can use any existing robust representation learning method, such as CC (Dai et al. 2022b), ISE (Zhang et al. 2022c), and LP (Lan et al. 2023). In this work, we build upon CC due to its simplicity. CC describes each cluster using a unique representation and maintains a dynamic cluster-level memory bank. In practice, it mainly involves three steps: 1) Initialize cluster memory bank; 2) conduct cluster-level contrastive learning; 3) update cluster memory bank. Next, we present these in detail.

Initialize Cluster Memory Bank. Given the calibrated features $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n\}$ and corresponding pseudo-labels $\{y_1, y_2, \ldots, y_n\}$, we initialize the cluster memory bank by calculating the mean feature vectors of each cluster, i.e.,

$$\mathbf{m}_i = \frac{1}{|\mathbb{C}_k|}\sum_{\mathbf{c}_i\in\mathbb{C}_k}\mathbf{c}_i, \qquad i = 1, 2, \ldots, k\,, \tag{8}$$

Algorithm 1: Improved clustering-based unsupervised object ReID with feature calibration

Require: Training dataset $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$.
for N in $(1, E_m)$ **do**
  Feature extraction by $\mathbf{F} = f_\theta(\mathbf{X})$
  Feature calibration by $\mathbf{C} = \mathcal{C}(\mathbf{F})$
  Pseudo-labels generation by DBSCAN
  Initialize cluster memory bank by Eq. (8).
  for $i = 1$ in $(1, iters)$ **do**
    Cluster-level contrastive learning by Eq.(9)
    Update cluster memory bank by Eq.(10)
  end for
end for

where $\mathbb{C}_k$ denotes k-th cluster, $|\mathbb{C}_k|$ is the number of samples in $\mathbb{C}_k$, and k is the number of clusters. Besides, $\mathbf{m}_i$ will be normalized to the unit hypersphere.

Cluster-Level Contrastive Learning. During training, we employ a random selection strategy, where we choose p person identities and extract z samples for each identity in every minibatch. Additionally, we regard the memory bank $\mathbf{M} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_k\}$ as a non-parametric classifier and use a ClusterNCE loss to train the model. The ClusterNCE loss is

$$\mathcal{L} = -\log \frac{\exp(\mathbf{c}_i \cdot \mathbf{m}^+/\tau)}{\sum_{j=1}^{k} \exp(\mathbf{c}_i \cdot \mathbf{m}_j/\tau)}, \tag{9}$$

where τ is a temperature hyper-parameter, $\mathbf{m}^+$ denotes the positive cluster representation of $\mathbf{c}_i$.

Update Cluster Memory Bank. Subsequently, we perform momentum updates to the cluster memory bank iteratively using the hard features extracted from the current minibatch, i.e.,

$$\mathbf{m}_i \leftarrow \alpha \mathbf{m}_i + (1-\alpha)\hat{\mathbf{c}}_i, \tag{10}$$

where α is the momentum parameter, $\hat{\mathbf{c}}_i$ denotes the hard features of i-th cluster, i.e.,

$$\hat{\mathbf{c}}_i = \arg\min_{\mathbf{c}_j \in \mathbb{C}_i} < \mathbf{c}_j, \mathbf{m}_i > . \tag{11}$$

where $< \bullet >$ denotes the inner product. Notably, the updated cluster vectors are also normalized to the unit hypersphere. In practice, we initialize the cluster memory bank per epoch while conducting cluster-level contrastive learning and updating the cluster memory bank per batch. A pseudo-code for improved clustering-based unsupervised object ReID with feature calibration is provided in Algorithm 1, where E_m denotes the number of epoch.

Remark. The proposed FCM can also be inserted into other clustering-based unsupervised object ReID tasks like unsupervised Visible-Infrared Person Re-Identification (Wu and Ye 2023) and Unsupervised Domain Adaptation (Li et al. 2023). One of the main changes is how to conduct robust representation learning.

Experiments

This section first provides the experimental details, then validates the significant benefits of FCM for clustering-based unsupervised object ReID, and highlights the impact on the representation space and clustering precision.

Experimental Details

Datasets The models are evaluated and compared on three widely-used benchmark object ReID datasets, which including Market-1501 (Zheng et al. 2015a) and MSMT17 (Wei et al. 2018), and DukeMTMC-reID (Zheng, Zheng, and Yang 2017). Market-1501 is collected by Tsinghua University, which has $32,668$ pedestrian images from 6 cameras, with a total of 1501 different pedestrian identities. The training set includes 751 people with $12,936$ images, and the remaining images are test set. MSMT17 comes from 15 cameras on campus, of which 12 cameras come from outdoors and 3 cameras come from indoors. It was based on Faster RCNN as a pedestrian detector and obtained $126,441$ images of 4101 pedestrians, of which $32,621$ images of 1041 pedestrians were used as the training set. It was collected from different time periods, which is closer to the real scene. The DukeMTMC-reID dataset is a large-scale pedestrian re-identification image dataset collected from 8 static cameras located on the Duke University campus. It comprises $36,411$ images, involving $1,404$ unique pedestrians. The training set consists of $16,522$ images from 702 different pedestrians, and the remaining portion forms the test set. It is noteworthy that the test set includes 408 pedestrians captured exclusively under a single camera, aimed at introducing some interfering factors.

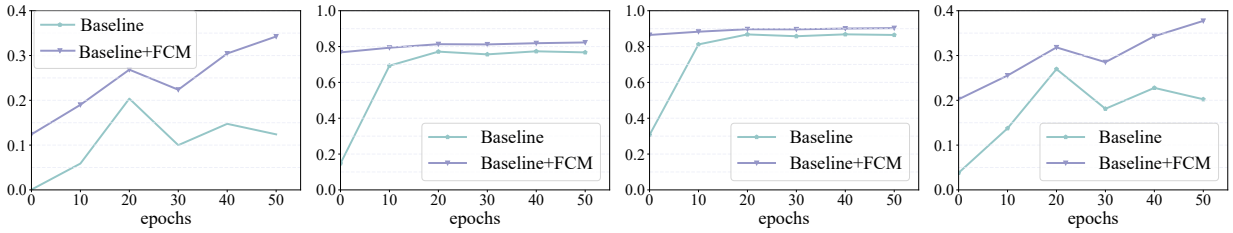
Metrics We follow the previous studies and use Cumulative Matching Characteristics (CMC) (Gray, Brennan, and Tao 2007) Rank-1, Rank-5, Rank-10, and mean Average Precision (mAP) (Zheng et al. 2015b) as the indicators for performance evaluation. Rank-k represents the fraction of queries that have at least one relevant sample in their k-nearest neighbors on a learned feature space.

Implementation Details While FCM can be adapted to most clustering-based unsupervised object ReID methods, in this study, we focus exclusively on a simple baseline, CC-ViT (He et al. 2021). This approach allows us to minimize the impact of external factors on the experimental results. In CC-ViT, the backbone employed is ViT (Dosovitskiy et al. 2020), which has been pre-trained on the large-scale unlabeled person ReID dataset LUPerson (Fu et al. 2021) using contrastive learning. We refer to the enhanced methods with our proposed FCM as *CC-FCM-ViT*. To ensure fair comparisons with the baseline, we maintain consistent experimental settings for *CC-FCM-ViT*.

The images of pedestrians are randomly cropped to 256×128. Then images undergo random flipping and erasing. The training epoch is 50 and the batch size is 256. The optimizer is Stochastic Gradient Descent (SGD). The learning rate is initialized to $3.5e-4$ and reduced by a factor of 10 every 20 epochs; the weight-decay of the optimizer is $5e-4$. The hyper-parameter ϵ of DBSCAN is fixed as 0.6, 0.8 and 0.7

Methods	Market-1501				MSMT17			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
MMCL (Wang and Zhang 2020)	45.5	80.3	89.4	92.3	11.2	35.4	44.8	49.8
HCT (Zeng et al. 2020b)	56.4	80.0	91.6	95.2	-	-	-	-
SpCL (Ge et al. 2020)	73.1	88.1	95.1	97.0	19.1	42.3	55.6	61,2
GCL (Chen et al. 2021)	66.8	87.3	93.5	95.5	21.3	45.7	58.6	64.5
IICS (Xuan and Zhang 2021)	72.9	89.5	95.2	97.0	26.9	56.4	68.8	73.4
JVTC+* (Chen et al. 2021)	75.4	90.5	96.2	97.1	29.7	54.4	68.2	74.2
ICE (Chen, Lagadec, and Bremond 2021)	82.3	93.8	97.6	98.4	38.9	70.2	80.5	84.4
OPLG-HCD (Zheng et al. 2021)	78.1	91.1	96.4	97.7	26.9	53.7	65.3	70.2
IIDS (Xuan and Zhang 2022)	78.0	91.2	96.2	97.7	35.1	64.4	76.2	80.5
HNGN (Li et al. 2022)	83.5	94.1	97.7	98.6	42.2	73.1	82.4	85.7
ISE (Zhang et al. 2022c)	85.3	94.3	98.0	98.8	37.0	67.6	77.5	81.0
PPLR (Cho et al. 2022)	84.4	94.3	97.8	98.6	42.2	73.3	83.5	86.5
RTMen (Yin et al. 2023)	83.1	93.9	97.7	98.4	40.8	72.0	81.5	84.6
FPM (Lan et al. 2023)	85.8	94.5	97.8	98.7	39.5	67.9	78.0	81.6
U-SSL (Wu et al. 2022)	82.3	94.1	97.4	98.8	43.1	73.2	89.4	90.8
CC-ViT	88.0	94.7	97.8	98.7	40.9	66.5	77.8	82.0
CC-FCM-ViT Ours	88.6	94.9	97.9	98.8	49.1	74.2	83.5	86.7
Δ	**(0.6 ↑)**	**(0.2↑)**	**(0.1↑)**	**(0.1↑)**	**(8.2↑)**	**(7.7↑)**	**(5.7↑)**	**(4.7↑)**
CC-ViT (with true labels)	90.4	96.0	98.6	99.1	64.6	83.7	91.7	93.9

Table 1: Comparison with the state-of-the-art methods on Market-1501 and MSMT17. ↑ indicates the improvement of our proposed method over baseline.



(a) ARI (b) AMI (c) VMS (d) FMS

Figure 5: Comparison of clustering quality across different epochs between baseline CC-ViT and CC-FCM-ViT.

on Market-1501, DukeMTMC-reID and MSMT17. The momentum parameter used in Eq. (10) is 0.2. Besides, we set the patch size $P = 16$ to divide the image into 16×8 blocks after passing through the convolution stem. We conduct all experiments using a single NVIDIA GeForce RTX 3090.

Experimental Results

We conduct a performance comparison between *CC-FCM-ViT* and several state-of-the-art methods on three object ReID datasets. The experimental results are presented in Table 1 and Table 2. It is worth noting that we have re-executed the baseline *CC-ViT* to ensure a fair comparison, and the results of residual methods are cited from prior studies. Furthermore, the 'with true labels' results are achieved by employing true labels in representation learning, representing the upper bounds of representation learning performance. Upon analyzing the experimental results, we identify two significant advantages of our method.

- *State-of-the-Art Results. CC-FCM-ViT* exhibits superior performance compared to other methods in most scenarios, achieving a substantial lead on the challenging MSMT17 dataset. While the improvements on the Market-1501 dataset may not be substantial, the performance of *CC-FCM-ViT* approaches that of the corresponding supervised method.
- *Significant Gains of Introducing FCM.* FCM demonstrates its effectiveness by consistently enhancing the performance of baseline methods across all datasets. Notably, the performance gains are particularly significant on the challenging MSMT17 and DukeMTMC-reID datasets. For instance, FCM boosts the mAP of *CC-ViT* by a notable 8.2% on MSMT17, providing strong validation for the efficacy of FCM.

In-Depth Analysis

Advantages Analysis of FCM As mentioned in the introduction, FCM has the capacity to refine the data distribution within the embedding space, thereby generating features conducive to clustering. To substantiate this claim, we present the t-SNE (Van Der Maaten 2014) visualization re-

Methods	DukeMTMC-reID			
	mAP	Rank-1	Rank-5	Rank-10
BUC (Lin et al. 2019)	27.5	47.4	62.6	68.4
HCT (Zeng et al. 2020b)	50.7	69.6	83.4	87.4
CAP (Wang et al. 2021)	67.3	81.1	89.3	91.8
IICS (Xuan and Zhang 2021)	64.4	80.0	89.0	91.6
RLCC (Zhang et al. 2021)	69.2	83.2	91.6	93.8
ICE (Chen, Lagadec, and Bremond 2021)	69.9	83.3	91.5	94.1
MGH (Wu et al. 2021)	70.2	83.7	92.1	93.7
MCRN (Wu et al. 2018)	69.9	83.5	-	-
MGCE-HCL (Sun, Li, and Li 2021)	67.5	82.5	-	-
PPLR (Cho et al. 2022)	68.4	82.5	90.4	92.9
ViT-small	69.6	81.3	88.8	91.3
CC-FCM-ViT Ours	71.9	83.8	91.0	93.2
Δ	**(2.3↑)**	**(2.5↑)**	**(2.2↑)**	**(1.9↑)**
CC-ViT (with true labels)	81.2	90.4	95.9	96.9

Table 2: Comparison with the state-of-the-art methods on DukeMTMC-reID. ↑ indicates the improvement of our proposed method over baseline.

Methods	CC-ViT			
	mAP	Rank-1	Rank-5	Rank-10
$k = 4$	47.2	72.8	82.2	85.7
$k = 8$	47.6	73.5	82.8	86.7
$k = 16$	49.1	74.2	83.5	86.7
$k = 32$	49.2	74.7	83.6	86.8
$k = 64$	48.0	72.8	82.4	86.1

Table 3: Investigation over the parameter k with $t = 1$.

l	MSMT17			
	mAP	Rank-1	Rank-5	Rank-10
$t = 0.05$	47.7	72.9	82.5	85.8
$t = 0.1$	48.0	73.4	82.7	86.1
$t = 0.5$	48.3	75.1	84.0	86.8
$t = 1$	49.1	74.2	83.5	86.7

Table 4: Investigation over the temperature t with $k = 32$.

sults in Figure 4 and provide a comparison of clustering precision in Figure 5. For the assessment of clustering quality, we employ four indicators (Pedregosa et al. 2011). The ARI values values range from $[-1, 1]$, while the other indicators vary within the range of $[0, 1]$.A higher score indicates a superior clustering result. Upon examining Figure 5, it becomes evident that FCM consistently enhances the quality of clustering during training. This effect stems from FCM's ability to encourage similar instances to converge in the feature space while facilitating the separation of dissimilar instances, as visually depicted in Figure 4.

Graph Structure of FCM To further validate the impact of graph structure on performance. In particular, there are two main parameters, k, the number of neighbors, t, the temperature value. In this test, we vary k and t in the sets $\mathcal{S}_k = \{4, 8, 16, 32, 64\}$ and $\mathcal{S}_t = \{0.05, 0.1, 0.5, 1\}$, respectively. Table 3 and Table 4 show that the performance is robust to parameter selection.

l	MSMT17			
	mAP	Rank-1	Rank-5	Rank-10
$layer = 1$	49.1	74.2	83.5	86.7
$layer = 2$	48.2	74.4	83.8	86.8
$layer = 3$	48.3	75.1	84.0	86.8
$layer = 5$	47.4	74.4	83.6	86.5

Table 5: Investigation over the number of SA Layers.

SA Layers of FCM This test validates the effect of SA layers in NpGAT. Specifically, we vary the number of layers l in the set $S_l = \{1, 2, 3, 5\}$. Table 5 shows that one SA layer is enough for generating sufficiently good pseudo-labels.

Conclusion

This work presents the first study exploring and demonstrating the efficacy of feature calibration for clustering-based unsupervised object ReID. We provide a simple yet effective feature calibration module. This module acts as a catalyst, enhancing the interaction between pseudo-label generation and representation learning. Extensive experiments on various object ReID benchmarks validate the effectiveness of our proposed feature calibration module. We hope that this work serves as an inspiration for future research endeavors, particularly in the direction of refining features before pseudo-label generation. In the future, we will validate the effectiveness of FCM in more unsupervised object ReID scenarios like unsupervised Visible-Infrared person ReID.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No.62276120 and No.62201453) and the Basic Research Project of Yunnan Province (Grant No. 202301AV070004).

References

Chen, H.; Lagadec, B.; and Bremond, F. 2021. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14960–14969.

Chen, H.; Wang, Y.; Lagadec, B.; Dantcheva, A.; and Bremond, F. 2021. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2004–2013.

Chen, Z.; Cui, Z.; Zhang, C.; Zhou, J.; and Liu, Y. 2023. Dual Clustering Co-teaching with Consistent Sample Mining for Unsupervised Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Cheng, D.; Tai, H.; Wang, N.; Wang, Z.; and Gao, X. 2022. Neighbour Consistency Guided Pseudo-Label Refinement for Unsupervised Person Re-Identification. *arXiv preprint arXiv:2211.16847*.

Cho, Y.; Kim, W. J.; Hong, S.; and Yoon, S.-E. 2022. Part-based pseudo label refinement for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7308–7318.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022a. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 1142–1160.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022b. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 1142–1160.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.

Fu, D.; Chen, D.; Bao, J.; Yang, H.; Yuan, L.; Zhang, L.; Li, H.; and Chen, D. 2021. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14750–14759.

Fu, Z.; Mao, Z.; Song, Y.; and Zhang, Y. 2023. Learning Semantic Relationship Among Instances for Image-Text Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15159–15168.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33: 11309–11321.

Gray, D.; Brennan, S.; and Tao, H. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, volume 3, 1–7.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15013–15022.

Hou, Z.; Yu, B.; and Tao, D. 2022. Batchformer: Learning to explore sample relationships for robust representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7256–7266.

Huang, J.; Zhao, F.; Zhou, M.; Xiao, J.; Zheng, N.; Zheng, K.; and Xiong, Z. 2023. Learning Sample Relationship for Exposure Correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9904–9913.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Jin, X.; He, T.; Shen, X.; Liu, T.; Wang, X.; Huang, J.; Chen, Z.; and Hua, X.-S. 2022. Meta clustering learning for large-scale unsupervised person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2163–2172.

Lan, L.; Teng, X.; Zhang, J.; Zhang, X.; and Tao, D. 2023. Learning to Purification for Unsupervised Person Re-identification. *IEEE Transactions on Image Processing*.

Li, D.; Wang, Z.; Wang, J.; Zhang, X.; Ding, E.; Wang, J.; and Zhang, Z. 2022. Self-Guided Hard Negative Generation for Unsupervised Person Re-Identification. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*.

Li, M.; Li, C.-G.; and Guo, J. 2022. Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 31: 3606–3617.

Li, S.; Li, F.; Li, J.; Li, H.; Zhang, B.; Tao, D.; and Gao, X. 2023. Logical Relation Inference and Multiview Information Interaction for Domain Adaptation Person Re-Identification. *IEEE Transactions on Neural Networks and Learning Systems*.

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, 8738–8745.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Seidenschwarz, J. D.; Elezi, I.; and Leal-Taixé, L. 2021. Learning intra-batch connections for deep metric learning. In *International Conference on Machine Learning*, 9410–9421. PMLR.

Sun, H.; Li, M.; and Li, C.-G. 2021. Hybrid contrastive learning with cluster ensemble for unsupervised person re-identification. In *Asian Conference on Pattern Recognition*, 532–546. Springer.

Van Der Maaten, L. 2014. Accelerating t-SNE using tree-based algorithms. *The journal of machine learning research*, 15(1): 3221–3245.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wang, D.; and Zhang, S. 2020. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10981–10990.

Wang, H.; Shen, J.; Liu, Y.; Gao, Y.; and Gavves, E. 2022. Nformer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7297–7307.

Wang, M.; Lai, B.; Huang, J.; Gong, X.; and Hua, X.-S. 2021. Camera-aware proxies for unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2764–2772.

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*, 79–88.

Wu, L.; Liu, D.; Zhang, W.; Chen, D.; Ge, Z.; Boussaid, F.; Bennamoun, M.; and Shen, J. 2022. Pseudo-pair based self-similarity learning for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 31: 4803–4816.

Wu, Y.; Wu, X.; Li, X.; and Tian, J. 2021. MGH: Metadata guided hypergraph modeling for unsupervised person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1571–1580.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Wu, Z.; and Ye, M. 2023. Unsupervised Visible-Infrared Person Re-Identification via Progressive Graph Matching and Alternate Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9548–9558.

Xuan, S.; and Zhang, S. 2021. Intra-inter camera similarity for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11926–11935.

Xuan, S.; and Zhang, S. 2022. Intra-inter domain similarity for unsupervised person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yan, T.; Zhu, K.; Zhu, G.; Tang, M.; Wang, J.; et al. 2022. Plug-and-Play Pseudo Label Correction Network for Unsupervised Person Re-identification. *arXiv preprint arXiv:2206.06607*.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 2872–2893.

Yin, J.; Zhang, X.; Ma, Z.; Guo, J.; and Liu, Y. 2023. A Real-Time Memory Updating Strategy for Unsupervised Person Re-Identification. *IEEE Transactions on Image Processing*.

Zeng, K.; Ning, M.; Wang, Y.; and Guo, Y. 2020a. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13657–13665.

Zeng, K.; Ning, M.; Wang, Y.; and Guo, Y. 2020b. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13657–13665.

Zhang, W.; Sheng, Z.; Yang, M.; Li, Y.; Shen, Y.; Yang, Z.; and Cui, B. 2022a. NAFS: A Simple yet Tough-to-beat Baseline for Graph Representation Learning. In *International Conference on Machine Learning*, 26467–26483. PMLR.

Zhang, W.; Yin, Z.; Sheng, Z.; Li, Y.; Ouyang, W.; Li, X.; Tao, Y.; Yang, Z.; and Cui, B. 2022b. Graph attention multi-layer perceptron. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4560–4570.

Zhang, X.; Ge, Y.; Qiao, Y.; and Li, H. 2021. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3436–3445.

Zhang, X.; Li, D.; Wang, Z.; Wang, J.; Ding, E.; Shi, J. Q.; Zhang, Z.; and Wang, J. 2022c. Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7369–7378.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Bu, J.; and Tian, Q. 2015a. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015b. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, 1116–1124.

Zheng, Y.; Tang, S.; Teng, G.; Ge, Y.; Liu, K.; Qin, J.; Qi, D.; and Chen, D. 2021. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8371–8381.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 3754–3762.