Discriminability Distillation in Group Representation Learning

Manyuan Zhang[1,2], Guanglu Song[1], Hang Zhou[2], and Yu Liu[1,2]⋆

[1] SenseTime X-Lab
[2] CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
{zhangmanyuan,songguanglu}@sensetime.com
zhouhang@link.cuhk.edu.hk, liuyuisanai@gmail.com

Abstract. Learning group representation is a commonly concerned issue in tasks where the basic unit is a group, set, or sequence. Previously, the research community tries to tackle it by aggregating the elements in a group based on an *indicator* either defined by humans such as the *quality* and *saliency*, or generated by a black box such as the attention score. This article provides a more essential and explicable view. We claim the most significant indicator to show whether the group representation can be benefited from one of its element is not the quality or an inexplicable score, but the *discriminability w.r.t.* the model. We explicitly design the *discrimiability* using embedded class centroids on a proxy set. We show the discrimiability knowledge has good properties that can be distilled by a light-weight distillation network and can be generalized on the unseen target set. The whole procedure is denoted as *discriminability distillation learning* (DDL). The proposed DDL can be flexibly plugged into many group-based recognition tasks without influencing the original training procedures. Comprehensive experiments on various tasks have proven the effectiveness of DDL for both accuracy and efficiency. Moreover, it pushes forward the state-of-the-art results on these tasks by an impressive margin.

Keywords: Group representation learning, set-to-set matching

1 Introduction

With the rapid development of deep learning and easy access to large-scale group data, recognition tasks using group information have drawn great attention in the computer vision community. The rich information provided by different elements can complement each other to boost the performance of tasks such as face recognition, person re-identification, and action recognition [54, 68, 19, 46, 62, 35, 42]. For example, recognizing a person through a sequence of frames is expected to be more accurate than watching only one image.

While traditional practice for group-based recognition is to either aggregate the whole set by average pooling [32, 49], max pooling [7], or just randomly

⋆ Corresponding author.

sampling [55], the fact that certain elements contribute negatively in recognition tasks has been ignored. Thus, the key problem for group-based recognition is how to define an efficient indicator to select representatives from sets.

To tackle such cases, previous methods aim at defining the "quality" or "saliency" for each element in a group [37, 62, 42, 40]. The weight for each element can be automatically learned by self-attention. For example, Liu et al. [37] propose the Quality Aware Network (QAN) to learn a quality score for each image inside an image set during network training. Other researchers adopt the same idea and extend it to specific tasks such as video-based person re-identification [33, 58] and action recognition [56] by learning spatial-temporal attentions. However, the whole quality/attention learning procedures are either manually designed or learned through a black box, which lacks explainability. Moreover, since previous attention and quality mechanism are mostly based on element feature, the features for all group elements need to be extracted, which is highly computational consuming.

In this work, we explore deeper into the underlying mechanism for defining effective elements. Assuming that a base network $\mathcal{M}$ has already been trained for element-based recognition using class labels, we define the "discriminability" of one element by how difficult it is for the network $\mathcal{M}$ to discriminate its class. How to measure the difficulty and the learning preference of the network $\mathcal{M}$ of elements remains an interesting problem. By considering the relationship between intra- and inter-class distance, we identify a successful discriminability indicator by *measuring one embedding's distance with all class centroids and compute the ratio of between positive and hardest-negative*. The *positive* is its distance from its class's corresponding centroid and the *hardest-negative* is its closest counterpart.

As the acquiring procedure of the discriminability indicator is highly flexible without either human supervision or network re-training, it can be adapted to any existing base. Though defined through trained bases, we find that the discriminability indicator can be easily distilled by training an additional light-weight network (Discriminability Distillation Network, DDNet). The DDNet takes the raw images as input and regresses the regularized discriminability indicators. We uniformly call the whole procedure *discriminability distillation learning* (DDL).

During inference, all elements are firstly sent into the light-weight DDNet to estimate their discriminability. Then element features will be weighted and aggregated according to their discriminability scores. In addition, in order to achieve the trade-off between accuracy and efficiency, we can filter elements by extracting and aggregating elements of high discriminability only. Since the base model tends to be heavy, the filtering process can save much computational cost. We evaluate the effectiveness of our proposed DDL on several classical yet challenging tasks including set-to-set face recognition, video-based person re-identification, and action recognition. Comprehensive experiments show the advantage of our method on both recognition accuracy and computational efficiency. State-of-the-art results can be achieved without modifying the base network.

We highlight our contributions as follows: (1) We define the *discriminability* of one element within a group from a more essential and explicable view, and propose an efficient indicator. Moreover, we demonstrate that the structure of discriminability distribution can be easily distilled by a light-weight network. (2) With a well-designed element discriminability learning and feature aggregating process, both efficiency and excellent performance can be achieved. We verify the good generalization ability of our discriminability distillation learning in many group-based recognition tasks, including set-to-set face recognition, video-based person re-identification, and action recognition through extensive studies.

2 Related work

Group representation learning which aims at formulating a unified representation has been proved efficient on various tasks [68, 37, 15, 55, 70]. In this paper, we care for three group representation learning tasks including set-to-set face recognition, video-based person re-identification, and action recognition. In this section, we will briefly review those related topics.

Set-to-Set Face Recognition. Set-to-set face recognition aims at performing face recognition [57, 27, 2, 29, 9, 69] using a set of images of a same person. To tackle set-to-set face recognition, traditional methods directly estimate the feature similarity among sets of feature vectors [1, 23, 5]. Other works seek to aggregate element features by simply applying max-pooling [7] or average pooling [32, 49] among set features to form a compact representation. However, since most set images are under unconstrained scenes, huge variations such as blur and occlusions will degrade the set feature discrimination. How to design a proper aggregation method for set face representation has been the key.

Recently, a few methods explore the manually defined operator or attention mechanism to form group representation. GhostVLAD [68] improves traditional VLAD. While Rao *et al.* [41] combine LSTM and reinforcement learning to discard low-quality element features. Liu *et al.* [37] and Yang *et al.* [62] introduce an attention mechanism to assign quality scores for different elements and aggregate feature vectors by quality weighted sum. To predict the quality score, an online attention network module is added and co-optimized by the target set-to-set recognition task. However, the definition of generated “quality” scores remains unclear and they are learned through a black box, which lacks explainability.

Video-Based Person Re-Identification. It is also beneficial to perform person re-identification [61, 39, 18, 16, 17, 33, 15] from videos. There are typically three components for video-based person re-identification: an image-level feature extractor, a temporal aggregating module, and the loss function [15]. Previous works mainly focus on optimizing the temporal aggregating module for video-based person re-identification. They can be divided into three categories, RNN-based [39, 61], attention-based [37, 71] and 3D-Conv based [15]. Yang *et al.* [61] model an RNN to encode element features and use the final hidden layer as the group feature representation. Liu *et al.* [37] use attention module to assign each element an quality score. While Gao *et al.* [15] directly utilize 3D Conv to encode the spatial-temporal feature for elements and propose a benchmark to compare different temporal aggregating module fairly.

Action Recognition. Action representation learning is another typical case of group-based representation learning. Real-world videos contain variable frames, so it is not practical to put the whole video to a memory limited GPU. The most usual approach for video understanding is to sample frames or clips and design late fusion strategies to form the video-level prediction.

Frame-based methods [64, 14, 46, 19] firstly extract frame features and aggregate them. Simonyan *et al.* [46] propose the two-stream network to simultaneously capture the appearance and motion information. Wang *et al.* [54] add attention module and learn to discard unrelated frames. Frame-based methods are computationally efficient, but only aggregating high-level frame features tends to limit the model's ability to handle complex motion.

Clip-based methods [50, 51, 13, 30] use 3D convolutional neural network to jointly capture spatial-temporal features. However, clip-based methods highly rely on the dense sample strategy, which introduces huge computational costs and makes it impractical to real-world applications. In this article, we show that by combining our DDL, the clip-based methods can achieve both excellent performance and computational efficiency.

3 Discriminability Distillation Learning

In this section, we first formulate the problem of group representation learning in section 3.1 and then define the discriminability in section 3.2. Next, we introduce the whole discriminability distillation learning (DDL) procedure in section 3.3. In sections 3.4 and 3.5, we discuss the aggregation method and the advantage of our DDL, respectively.

3.1 Formulation of Group Representation Learning

Compared to using a single element, performing recognition with group representation can further explore the complementary information among group elements and benefit from them. For example, recognizing a person from a group of his photos instead of one image is sure to facilitate the result.

The most popular way to handle group-based recognition tasks is to formulate a unified representation for a whole group of elements [37, 68, 55, 15]. Suppose a base network $\mathcal{M}$ is trained for the element-based recognition task. Define $f_i \in \mathbb{R}^d$ as the embedded feature of element I_i in group $\mathbf{I}_S$ from $\mathcal{M}$, the unified feature representation of the whole group is

$$f_{\mathbf{I}_S} = \mathcal{G}(f_1, f_2, \cdots, f_i), \tag{1}$$

where $\mathcal{G}$ indicates the feature aggregation module. While previous research has revealed that conducting $\mathcal{G}$ with quality [37] has priority over simple aggregation, this kind of method is not explainable and computation-consuming. In this

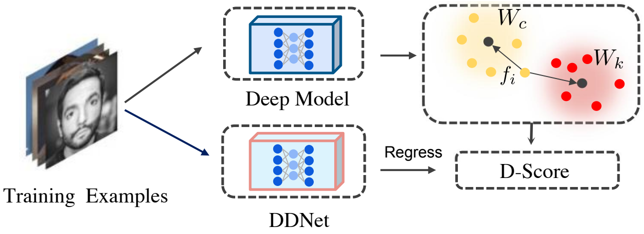

Fig. 1. The pipeline of group representation learning with DDL. Given a base feature extracting model, we first compute the discriminability for each training element and then train a light-weight discriminability distillation network (DDNet) to regress it. The discriminability is formulated from the view of intra and inter-class distance with class centroids for element

article, we propose discriminability distillation learning (DDL) to generate the *discriminability* of feature representation.

3.2 Formulation of Discriminability

Towards learning efficient and accurate $\mathcal{G}$, we propose to define the *discriminability* of elements to replace the traditional quality or attention mechanism.

After training the base model $\mathcal{M}$ on the classification task, features of the training elements from the same class are projected to hyperspace tightly in order to form an implicit decision boundary and minimizing target loss [36]. This statement exists when $\mathcal{M}$ is supervised by all kinds of loss functions (softmax-cross entropy [47], triplet [44] or margin-based [9, 12] losses). Our key observation is that the features embedded close to their corresponding class centroids are normally the representative examples, while features far away or closer to other centroids are usually the confusing ones.

Based on our motivation, we jointly consider the feature space distribution and explicitly distill the *discriminability* by encoding the intra-class distance and inter-class distance with class centroids. Let $\mathcal{X}$ denotes the training set with K classes and $C_m \in \mathbb{R}^d, m \in [1, K]$ is the class centroid of class m, which is the average of features. For feature $f_i, i \in [1, s]$ where s denotes the size of $\mathcal{X}$. Assume the positive class for f_i is p, while the negatives are $n \in [1, K], n \neq p$. The intra-class distance and inter-class distance for f_i are formulated as:

$$\begin{aligned} dist_{ip} &= \frac{f_i \cdot C_p}{\|f_i\|_2 \; \|C_p\|_2}, \\ dist_{in} &= \frac{f_i \cdot C_n}{\|f_i\|_2 \; \|C_n\|_2}, \;\; n \in [1, K], n \neq p. \end{aligned} \tag{2}$$

Here we use the cosine distance as feature distance metric. Other metrics like Euclidean distance are also applicable. Then the *discriminability* $\mathcal{D}_i$ of f_i can be defined as:

$$\mathcal{D}_i = \frac{dist_{ip}}{\max\left\{dist_{in} \mid n \in [1,K], n \neq p\right\}}. \tag{3}$$

It is the ratio between the feature's distance from the centroid of its own class and the distance from the hardest-negative class. Considering the variant number of elements in different groups, we further normalize the *discriminability* by:

$$\mathcal{D}_i = \tau\left(\frac{\mathcal{D}_i - \mu(\{\mathcal{D}_j \mid j \in [1,s]\})}{\sigma(\{\mathcal{D}_j \mid j \in [1,s]\})}\right) \tag{4}$$

where $\tau(\cdot)$, $\mu(\cdot)$ and $\sigma(\cdot)$ denote the sigmoid function, the mean value and the standard deviation value of $\{\mathcal{D}_j \mid j \in [1,s]\}$, respectively. We denote the normalized $\mathcal{D}_i$ as discriminability score (D-score).

Cooperated with the feature space distribution, the discriminability $\mathcal{D}_i$ is more interpretable and reasonable. It can discriminate features better by explicitly encoding the intra- and inter-class distances with class centroids.

3.3 Discriminability Distillation Learning

From section 3.2, given a base model $\mathcal{M}$ and its training dataset, the $\mathcal{D}_i$ of f_i can be naturally computed by Eq (2)-(4). However, the score is unavailable to test set $\mathcal{T}$. In order to estimate unseen element's discriminability, we formulate the discriminability distillation learning (DDL) procedure for group representation.

Our idea is to *distill* the discriminability explicitly using a light-weight auxiliary network from the training samples. It is called the Discriminability Distillation Network (DDNet). Denote the DDNet as $\mathcal{N}$, the approximated $\hat{\mathcal{D}}_i$ for $\mathcal{D}_i$ can be given by:

$$\hat{\mathcal{D}}_i = \mathcal{N}(I_i; \boldsymbol{\theta}), \tag{5}$$

where $\boldsymbol{\theta}$ denotes the parameters of $\mathcal{N}$. To train $\mathcal{N}$, we apply mean squared error between $\hat{\mathcal{D}}_i$ and target $\mathcal{D}_i$ as

$$L = \frac{1}{2N}\sum_{i}^{N}(\hat{\mathcal{D}}_i - \mathcal{D}_i)^2, \tag{6}$$

where N is the batch size. The training is conducted with the same training set for the base model and there is no need to modify the base model $\mathcal{M}$.

3.4 Feature Aggregation $\mathcal{G}$

During inference, we can generate $\hat{\mathcal{D}}_i$ via Eq (5) for each element I_i in the given element set I_S. Then we can filter out some elements with low discriminability in order to accelerate the feature extracting process of $\mathcal{M}$. Given the pre-defined threshold t and base model $\mathcal{M}$, the group element feature extracting process is

$$f_i = \mathcal{M}(I_i), \hat{\mathcal{D}}_i > t, \tag{7}$$

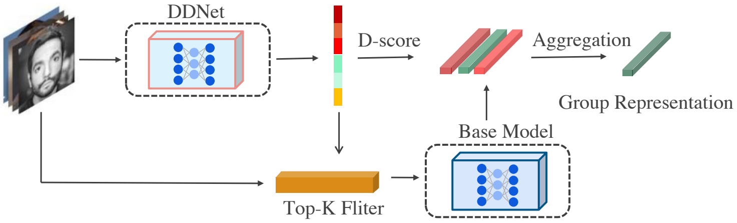

Fig. 2. The pipeline of the test stage with DDL. For a group of elements, we first predict D-score by the trained-well DDNet for each element. Then we will filter elements by their D-scores and only extract feature for those elements with high D-scores by the base model. Finally extracted features will be weighted sum to form the group representation

and $\mathcal{G}$ in Eq (1) can be formulated as:

$$F_{\mathbf{I}_S} = \mathcal{G}(f_1, f_2, \cdots, f_n) = \sum_{i}^{n} \frac{\hat{\mathcal{R}}_i f_i}{\hat{\mathcal{R}}_i}, \tag{8}$$

where n is the number of I_S whose discriminability is higher than threshold t, and $\hat{\mathcal{R}}_i$ is the re-scaled D-score via

$$\hat{\mathcal{R}}_i = K\hat{\mathcal{D}}_i + B. \tag{9}$$

In Eq (9), we scale the D-score of element set I_S between 0 and 1 to ensure the same range for element sets with different lengths. K and B are formulated as

$$K = \frac{1}{\max\{\hat{\mathcal{D}}_i \mid i \in [1,n]\} - \min\{\hat{\mathcal{D}}_i \mid i \in [1,n]\}}, \tag{10}$$

$$B = 1 - K \max\{\hat{\mathcal{D}}_i \mid i \in [1,n]\}. \tag{11}$$

3.5 Advantage of Discriminability Distillation Learning

Different from the subjective quality judgment of an image or the attention mechanism, we explicitly assign *discriminability* for each element via the feature space distribution. By jointly considering the inter- and intra-class distances with class centroids, DDL can effectively approximate how discriminative a feature is. By aggregating more information with features with high *discriminability*, more discriminative group representation can be formed, leading to a significant performance boost for group-based recognition tasks. In addition, the well-design discriminability distillation learning process needn't modify the base model, making it easy to be plugged into many popular recognition frameworks. Furthermore, We can change the threshold for the discriminability filtering process according to different application scenarios to achieve a trade-off between accuracy and computational cost.

4 Experiments

We evaluate our DDL on three popular group-based recognition tasks: set-to-set face recognition, video-based person re-identification, and action recognition. An ablation study will be conducted along with the set-to-set face recognition experiments.

4.1 Set-to-Set Face Recognition

In this section, we evaluate DDL for set-to-set face recognition on four datasets including two video datasets: YouTube Face (YTF) [57], iQIYI-VID-FACE [26]; and two template-based datasets: IARPA Janus Benchmark A (IJB-A) [29] and IARPA Janus Benchmark C (IJB-C).

4.1.1 Implementation Details. For data pre-processing, RetinaFace [11] is used to detect faces and their corresponding landmarks for all datasets. Images are aligned to 112 × 112 by similarity transformation with facial landmarks.

We train our base model and DDNet on the MS-Celeb-1M dataset [21] cleaned by [9]. The base model we select is modified ResNet-101 [24] released by [9]. As for the DDNet, we use a light-weight channel reduced ResNet-18 network, whose channels for 4 stages are $\{8, 16, 32, 48\}$, respectively. It only introduces 81.9 Mflops, which is super-efficient.

The loss function for the base model training is ArcFace [9] and the total training step is 180k with initial learning rate 0.1 on 8 NVIDIA Tesla V100 GPUs. The training process for our DDNet is similar to the base model. The default discriminability threshold we select is 0.15, empirically.

4.1.2 Evaluation on YouTube Face. The YouTube Face [57] dataset includes 3425 videos of 1595 identities with an average of 2.15 videos per identity. The videos vary from 48 frames to 6,070 frames. We report the 1:1 face verification accuracy of the given 5,000 video pairs in our experiments.

As shown in Table 1, our DDL achieves state-of-the-art performance on the YouTube Face benchmark [57]. It outperforms [9] by 0.16% and other set-to-set face recognition methods by impressive margins. For comparison with different aggregation strategies like average pooling, DDL can boost performance by 0.21%, which indicates DDL has learned a meaningful pattern for discriminability. As a post-training module, DDL can cooperate with any existing base. Note that if we only select the top-1 discriminability frame, DDL can also achieve 97.08%, which achieves above 130x acceleration. The computation complexity for the base model is 11 Gflops (ResNet-101) while our DDNet only introduces 81.9 Mflops. By filtering most frames, great computational cost is saved.

4.1.3 Evaluation on IQIYI-VID-FACE. Since the results on YouTube Face benchmark tend to be saturated, we test our DDL on the challenging

Table 1. Video face verification performance on YouTube Face dataset, compared with state-of-the-art methods and baseline methods

Method	Accuracy(%)	Method	Accuracy(%)
Li *et al.* [32]	84.8	DeepFace [49]	91.4
FaceNet [44]	95.52	NAN [62]	95.72
DeepID2 [48]	93.20	QAN [37]	96.17
C-FAN [20]	96.50	Rao *et al.* [42]	96.52
Liu *et al.* [38]	96.21	Rao *et al.* [41]	94.28
CosFace [53]	97.65	ArcFace [9]	98.02
Average	97.97	*Top 1*	97.08
		DDL	**98.18**

Table 2. Comparison with different participants and aggregation strategy on the IQIYI-VID-FACE challenge. By combining with PolyNet, DDL achieves state-of-the-art performance

Method	TPR@FPR=1e-4(%)	Method	TPR@FPR=1e-4(%)
MSRA	71.59	Alibaba-VAG	71.10
Insightface	67.00	DDL (PolyNet)	**72.98**
Average	65.84	*Top 1*	65.22
DDL w/o re-scale	67.38	DDL	**69.05**

video face verification benchmark IQIYI-VID-FACE [10], The IQIYI-VID-FACE dataset aims to identify the person in entertainment videos by face images. It is the largest video face recognition test benchmark so far, containing 643,816 video clips of 10,034 identities. The test protocol is 1:1 verification, and the True Accept Rate (TAR) under False Accept Rate (FAR) at 1e-4 is reported.

As shown in Table 2, compared with the average pooling, DDL improves performance by 3.21%. Even only aggregating the top-1 discriminability score frame can still achieve an equal performance of average aggregation for all frames. It shows that our DDL has selected the most discriminative element of the set. By combining stronger base model PolyNet [65], our DDL achieves state-of-the-art performance on the IQIYI-VID-FACE challenge.

4.1.4 Evaluation on IJB-A and IJB-C.

The IARPA Janus Benchmark A (IJB-A) [29] and IARPA Janus Benchmark C (IJB-C) are unconstrained face recognition benchmarks. They are template-based test benchmarks where both still images and video frames are included in templates. IJB-A containing 25, 813 faces images of 500 identities while IJB-C has 140, 740 faces images of 3, 531 subjects. Since the images in IJB-C dataset have large variations, it is regarded as a challenging set-to-set face recognition benchmark.

Table 3. Performance comparisons on IJB-A verification benchmark. The True Accept Rates (TAR) vs. False Accept Rate (FAR) are reported

Method	IJB-A (TAR@FAR)		
	FAR=1e-3(%)	FAR=1e-2(%)	FAR=1e-1(%)
Template Adaptation [8]	83.6 ± 2.7	93.9 ± 1.3	97.9 ± 0.4
TPE [43]	81.30 ± 2.0	91.0 ± 1.0	96.4 ± 0.5
Multicolumn [60]	92.0 ± 1.3	96.2 ± 0.5	98.9 ± 0.2
QAN [37]	89.31 ± 3.92	94.2 ± 1.53	98.02 ± 0.55
VGGFace2 [4]	92.1 ± 1.4	96.8 ± 0.6	99 ± 0.2
NAN [62]	88.1 ± 1.1	94.1 ± 0.8	97.8 ± 0.3
GhostVLAD [68]	93.5 ± 1.5	97.2 ± 0.3	99.0 ± 0.2
Liu *et al.* [38]	93.61 ± 1.51	97.28 ± 0.28	98.94 ± 0.31
ArcFace [9]	97.89 ± 1.5	98.51 ± 0.3	99.05 ± 0.2
Average	97.71 ± 0.6	98.43± 0.4	99.01± 0.2
DDL	**98.44 ± 0.3**	**98.79 ± 0.2**	**99.13 ± 0.1**

Table 4. Performance comparisons on IJB-C verification benchmark. The True Accept Rates (TAR) vs. False Accept Rate (FAR) are reported

Method	IJB-C (TAR@FAR)				
	1e-6(%)	1e-5(%)	1e-4(%)	1e-3(%)	1e-2(%)
Yin *et al.* [63]	-	-	-	0.1	83.8
Xie *et al.* [59]	-	-	88.5	94.7	98.3
Zhao *et al.* [66]	-	82.6	89.5	93.5	96.2
multicolumn [60]	-	77.1	86.2	92.7	96.8
VGGFace2 [4]	-	74.7	84.0	91.0	96.0
PFE [45]	-	89.64	93.25	95.49	97.17
ArcFace [9]	86.25	93.15	95.65	97.20	98.18
Average	86.69	92.72	94.89	96.62	97.90
DDL	**92.39**	**94.89**	**96.41**	**97.47**	**98.33**

Tables 3 and 4 show the results on the IJB-A and IJB-C benchmark for different methods. From the two tables, we can see that our DDL improves verification performance by a convincing margin with average pooling for both two benchmarks, especially under severe FAR at 1e-5 by 0.73% on IJB-A and FAR at 1e-6 by 5.7% on IJB-C. Compared with IJB-A, IJB-C has more images and covers more variations among images, such as pose, blur, resolution, and conditions. So the performance gain with DDL is larger.

Compared with the state-of-the-art methods, our DDL improves IJB-A by 0.55% when FAR =1e-3 and IJB-C by 6.14% when FAR = 1e-6. These results indicate the effectiveness and robustness of our DDL. What's more, unlike many previous methods that need fine-tune with the base model on set-to-set recognition training datasets [37, 68], the only supervision for DDL training is the discriminability generated with the base model on the same training set, which is highly flexible.

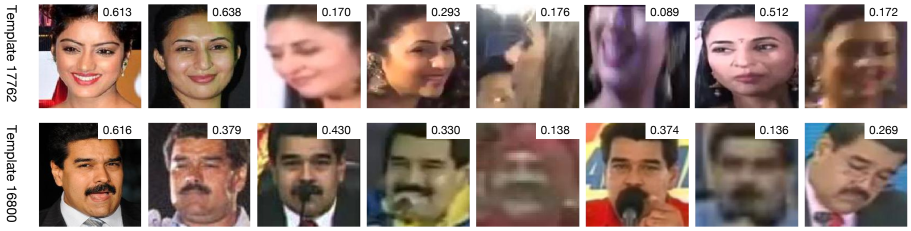

Fig. 3. The visualization results of discriminability for images of Template ID 17762 and 16800 from IJB-C dataset

To qualitatively evaluate the discriminability pattern learned by our DDL, we visualize the discriminability score distribution for two template images in IJB-C datasets. As shown in Figure 3, DDL can effectively identify image discriminability. Images with large poses, visual blur, occlusion, and incomplete content are regarded to be low discriminative. The efficient discriminability judgment ability for our DDL leads to an extraordinary performance on set-to-set face recognition problems.

4.1.5 Ablation Study

The architecture of DDNet and base model. In the above experiments, we have adopted the channel reduced version of ResNet-18 as the backbone for DDNet. When inference, all test samples will be sent to DDNet firstly to predict discriminability. Therefore, the test computational cost is very sensitive to the architecture of DDNet. We design it as light-weight as possible. We also conduct experiments with wider and deeper DDNet and test on IJB-C. As shown in Table 5, the wider and deeper networks have not brought significant performance gains.

As for the base model, we also experiment DDL with MobileFaceNet [6], a popular backbone for mobile devices. From Table 5, we can see that by combining will DDL, a consistent performance gain can be achieved on set-to-set face recognition task for MobileFaceNet.

Train on other datasets. In the aforementioned experiments, we use the MS-Celeb-1M dataset for the base model and DDNet training. To demonstrate the good generalization of our method, we also train the base model and DDNet with IMDB-Face [52] dataset. IMDb-Face is a new large-scale noise-controlled dataset for face recognition. The dataset contains about 1.7 million faces, 59k identities, which is manually cleaned from 2.0 million raw images. The results on IJB-C are shown in Table 5, DDL improves set-to-set face recognition by a huge margin compared with simple average pooling, up to 15.23% at FPR=1e-6. The model trained on IMDB-Face tends to be weaker than MS-Celeb-1M and more

Table 5. Ablation study with different DDNet architecture, base model architecture and training datasets. Results are reported on IJB-C benchmark. 'CD' means channel reduced

Method			IJB-C (TAR@FAR)				
DDNet	Base Model	Train Datasets	1e-6	1e-5	1e-4	1e-3	1e-2
ResNet-18-CD	ResNet-101	MS-Celeb-1M	**91.14**	**95.75**	**96.94**	97.72	**98.36**
ResNet-34-CD	ResNet-101	MS-Celeb-1M	91.13	95.74	96.90	97.72	98.33
ResNet-18	ResNet-101	MS-Celeb-1M	90.93	95.74	96.92	**97.73**	98.35
ResNet-18-CD	MobileFaceNet	MS-Celeb-1M	**87.32**	**91.45**	**94.30**	**96.24**	**97.82**
-	MobileFaceNet	MS-Celeb-1M	79.88	88.21	92.08	95.22	97.24
ResNet-18-CD	ResNet-101	IMDB-Face	**88.35**	**92.26**	**95.09**	**96.71**	**98.05**
-	ResNet-101	IMDB-Face	73.12	86.44	92.44	94.70	97.40

Table 6. Ablation study with loss function. Results are reported on YouTube Face benchmark

Method		Accuracy(%)
DDL	loss function	
✓	ArcFace	**98.18**
×	ArcFace	97.97
✓	CosFace	**97.91**
×	CosFace	97.68
✓	SphereFace	**97.12**
×	SphereFace	96.83

easily confused by hard negative pairs, thus DDL achieves a more significant improvement.

The influence of re-scale. In Eq (9), we re-scale the discriminability scores of element set between 0 and 1 to ensure the same range for element sets with different lengths. In this part, we compare the re-scale strategy and origin scale on IQIYI-VID-FACE benchmark. As shown in Table 2, re-scale can boost performance for 1.67%. For video face recognition, which contains various frames from dozens of to thousands of, it is necessary to re-scale the predicted discriminability scores at the test stage.

Combined with more loss functions. There are many successful loss function these years for face recognition task, such as ArcFace [9], CosFace [53] and SphereFace [34]. We combine DDL with more loss functions and test on YouTube Face benchmark. As shown in Table 6, all loss functions achieve constant performance gain with DDL. DDL is not sensitive to the base model training loss function and can easily cooperate with any existing base.

4.2 Video-Based Person Re-Identification

In this section, we will evaluate our DDL with the video-based person re-identification task on Mars [67]. It the largest video-based person re-identification dataset. The train and test set are followed official split.

Table 7. Results for video-based person re-identification on Mars

	mAP	CMC-1	CMC-5	CMC-20
Zheng *et al.*[67]	45.6	65.0	81.1	88.9
Li *et al.* [31]	56.1	71.8	86.6	93.1
QAN [37]	51.7	73.7	84.9	91.6
Hermans *et al.*[25]	67.7	79.8	91.4	-
3D conv [15]	70.5	78.5	90.9	95.9
Atttention [15]	76.7	83.3	93.8	97.4
RNN [15]	73.9	81.6	92.8	96.3
average	74.1	81.3	92.6	96.7
DDL	**77.7**	**84.0**	**94.8**	**97.4**

To train the base model, triplet loss function and softmax cross-entropy loss function are used. The similarity metric is L2 distance. Standard ResNet-50 pre-trained on ImageNet is used and video frames are resized to 224×112. We will report mean average precision score (mAP) and cumulative matching curve (CMC) at rank-1, rank-5 and rank-20. Note that re-rank is not applied in the comparison.

The results are shown in Table 7, DDL boosts the performance consistently. Compared with average pooling, DDL achieves performance gain for 3.6% mAP. For more complicated aggregation strategies like RNN and the state-of-the-art attention mechanism, DDL also improves performance. The good performance of video-based person re-identification further demonstrates the efficiency of our DDL in group representation learning.

4.3 Action Recognition

In this section, we will evaluate our DDL on two most popular action recognition datasets ActivityNet-1.2 [3] and Kinetics-700 [28]. The ActivityNet-1.2 contains 4,819 training videos and 2,383 validation videos for 100 action class. It is an untrimmed video dataset, namely more temporal variance and noises there are. The Kinetics-700 is a well-trimmed action recognition datasets, which contains over 650k videos from 700 classes.

All video frames are extracted by FFmpeg with 30fps then resized and center crop to 112×112. We select three clip-based action recognition baseline method, the 3D-ResNet-50 [22], SlowFast-50 [13] and R(2+1)D-50 [51]. The training config for those base models follows SlowFast [13]. In the original approach, all three methods rely on dense sampling during testing. To be more specific, they oversampling both spatially and temporally to capture target activation.

The DDNet architecture for action recognition is the same with image task, but replace all 2D-Conv to 3D-Conv. A video will firstly be divided into many clips, and each clip's discriminability will be generated by DDNet, only top-K clips will be extracted feature and aggregated. The K we select for ActivityNet-1.2 and Kinetics is 9 and 5, respectively. We select random and uniform sampling

Table 8. Video action recognition results(%) on ActivityNet-1.2 dataset. Accuracy is reported by top-1 on the validation set

Model	DDL	Random	Uniform	Dense
clip number		9		60
3D-RS-50	**86.38**	82.83	83.14	83.92
R(2+1)D-RS-50	**89.08**	84.51	84.89	85.46
SlowFast-RS-50	**90.21**	85.92	86.14	87.72

Table 9. Video action recognition results(%) on Kinetics-700 dataset. Accuracy is reported on the validation set and is the average of top1 and top 5 accuracy

Model	DDL	Random	Uniform	Dense
clip number		5		30
3D-RS-50	**71.01**	68.26	67.43	68.83
R(2+1)D-RS-50	**72.51**	69.24	68.79	70.94
SlowFast-RS-50	**74.23**	72.39	72.05	73.77

K clips for comparison with sampling by DDL. A dense sampling experiment is also conducted.

From Table 8, DDL improves recognition performance for all baseline models on ActivityNet-1.2. For the state-of-the-art clips-based model SlowFast, combining it with DDL can achieve around 4% accuracy gain compared with random or uniform sampling on ActivityNet-1.2. What's more, DDL can even outperform dense sampling by 2.49%, while the dense sampling strategy sample above 5x more clips (estimated by the average duration 120s for ActivitNet-1.2).

For Kinetics-700, the results are in Table 9. DDL outperforms random sampling by 1.84% and uniform sampling by 2.18%. For dense sampling, DDL can achieve 0.46% gain with 6x speed up. Since the Kinetics-700 is trimmed by human and video quality is under control, combining with DDL can also significantly boost recognition performance and save computational consumption.

5 Conclusion

In this paper, we have proposed a novel post-processing module called Discriminability Distillation Learning (DDL) for all group-based recognition tasks. We explicitly define the discriminability with observations on feature embedding, then apply a light-weight network for discriminability distillation and feature aggregation. We identify the advantage of our proposed methods in the following aspects: (1) The entire discriminability distillation is performed without modifying the pre-trained based network, which is highly flexible comparing with existing quality-aware or attention methods. (2) Our distillation network is extremely light-weighted which saves great computational cost. (3) With our DDL and feature aggregation, we achieve state-of-the-art results on multiple group-based recognition tasks including set-to-set face recognition, video-based person re-identification, and action recognition.

References

1. Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face recognition with image sets using manifold density divergence. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 581–588. IEEE (2005)
2. Beveridge, J.R., Phillips, P.J., Bolme, D.S., Draper, B.A., Givens, G.H., Lui, Y.M., Teli, M.N., Zhang, H., Scruggs, W.T., Bowyer, K.W., et al.: The challenge of face recognition from digital point-and-shoot cameras. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). pp. 1–8. IEEE (2013)
3. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE (2018)
5. Cevikalp, H., Triggs, B.: Face recognition based on image sets. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2567–2573. IEEE (2010)
6. Chen, S., Liu, Y., Gao, X., Han, Z.: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition. pp. 428–438. Springer (2018)
7. Chowdhury, A.R., Lin, T.Y., Maji, S., Learned-Miller, E.: One-to-many face recognition with bilinear cnns. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2016)
8. Crosswhite, N., Byrne, J., Stauffer, C., Parkhi, O., Cao, Q., Zisserman, A.: Template adaptation for face verification and identification. Image and Vision Computing **79**, 35–48 (2018)
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
10. Deng, J., Guo, J., Zhang, D., Deng, Y., Lu, X., Shi, S.: Lightweight face recognition challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
11. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641 (2019)
12. Duan, Y., Lu, J., Zhou, J.: Uniformface: Learning deep equidistributed representation for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3415–3424 (2019)
13. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. arXiv preprint arXiv:1812.03982 (2018)
14. Gan, C., Gong, B., Liu, K., Su, H., Guibas, L.J.: Geometry guided convolutional neural networks for self-supervised video representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5589–5597 (2018)
15. Gao, J., Nevatia, R.: Revisiting temporal modeling for video-based person reid. arXiv preprint arXiv:1805.02104 (2018)

16. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. ICLR (2020)
17. Ge, Y., Chen, D., Zhu, F., Zhao, R., Li, H.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. arXiv preprint arXiv:2006.02713 (2020)
18. Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al.: Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In: Advances in neural information processing systems. pp. 1222–1233 (2018)
19. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: Actionvlad: Learning spatio-temporal aggregation for action classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 971–980 (2017)
20. Gong, S., Shi, Y., Jain, A.K.: Video face recognition: Component-wise feature aggregation network (c-fan). arXiv preprint arXiv:1902.07327 (2019)
21. 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. pp. 87–102. Springer (2016)
22. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)
23. Harandi, M.T., Sanderson, C., Shirazi, S., Lovell, B.C.: Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In: CVPR 2011. pp. 2705–2712. IEEE (2011)
24. 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. pp. 770–778 (2016)
25. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
26. iQIYI: iqiyi-vid-face (2019), http://challenge.ai.iqiyi.com/data-cluster
27. Kalka, N.D., Maze, B., Duncan, J.A., O'Connor, K., Elliott, S., Hebert, K., Bryan, J., Jain, A.K.: Ijb–s: Iarpa janus surveillance video benchmark. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–9. IEEE (2018)
28. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
29. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1931–1939 (2015)
30. Korbar, B., Tran, D., Torresani, L.: Scsampler: Sampling salient clips from video for efficient action recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6232–6242 (2019)
31. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 384–393 (2017)
32. Li, H., Hua, G., Shen, X., Lin, Z., Brandt, J.: Eigen-pep for video face recognition. In: Asian conference on computer vision. pp. 17–33. Springer (2014)
33. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)

34. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
35. Liu, X., Guo, Z., Li, S., Kong, L., Jia, P., You, J., Kumar, B.: Permutation-invariant feature restructuring for correlation-aware image set-based recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4986–4996 (2019)
36. Liu, Y., Song, G., Shao, J., Jin, X., Wang, X.: Transductive centroid projection for semi-supervised large-scale recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 70–86 (2018)
37. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5790–5799 (2017)
38. Liu, Z., Hu, H., Bai, J., Li, S., Lian, S.: Feature aggregation network for video face recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
39. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1325–1334 (2016)
40. Nikitin, M.Y., Konouchine, V.S., Konouchine, A.: Neural network model for video-based face recognition with frames quality assessment. Computer Optics **41**(5), 732–742 (2017)
41. Rao, Y., Lin, J., Lu, J., Zhou, J.: Learning discriminative aggregation network for video-based face recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3781–3790 (2017)
42. Rao, Y., Lu, J., Zhou, J.: Attention-aware deep reinforcement learning for video face recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3931–3940 (2017)
43. Sankaranarayanan, S., Alavi, A., Castillo, C.D., Chellappa, R.: Triplet probabilistic embedding for face verification and clustering. In: 2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS). pp. 1–8. IEEE (2016)
44. 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. pp. 815–823 (2015)
45. Shi, Y., Jain, A.K.: Probabilistic face embeddings (2019)
46. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
47. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems. pp. 1988–1996 (2014)
48. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2892–2900 (2015)
49. 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. pp. 1701–1708 (2014)
50. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)

51. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
52. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Change Loy, C.: The devil of face recognition is in the noise. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 765–780 (2018)
53. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
54. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4325–4334 (2017)
55. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
56. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
57. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. IEEE (2011)
58. Wu, L., Wang, Y., Gao, J., Li, X.: Where-and-when to look: Deep siamese attention networks for video-based person re-identification. IEEE Transactions on Multimedia **21**(6), 1412–1424 (2018)
59. Xie, W., Shen, L., Zisserman, A.: Comparator networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 782–797 (2018)
60. Xie, W., Zisserman, A.: Multicolumn networks for face recognition. arXiv preprint arXiv:1807.09192 (2018)
61. Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X.: Person re-identification via recurrent feature aggregation. In: European Conference on Computer Vision. pp. 701–716. Springer (2016)
62. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4362–4371 (2017)
63. Yin, B., Tran, L., Li, H., Shen, X., Liu, X.: Towards interpretable face recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9348–9357 (2019)
64. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4694–4702 (2015)
65. Zhang, X., Li, Z., Change Loy, C., Lin, D.: Polynet: A pursuit of structural diversity in very deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 718–726 (2017)
66. Zhao, J., Cheng, Y., Cheng, Y., Yang, Y., Zhao, F., Li, J., Liu, H., Yan, S., Feng, J.: Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9251–9258 (2019)
67. 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. pp. 868–884. Springer (2016)

68. Zhong, Y., Arandjelović, R., Zisserman, A.: Ghostvlad for set-based face recognition. In: Asian Conference on Computer Vision. pp. 35–50. Springer (2018)
69. Zhou, H., Liu, J., Liu, Z., Liu, Y., Wang, X.: Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
70. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: AAAI Conference on Artificial Intelligence (AAAI) (2019)
71. 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 reidentification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4747–4756 (2017)