Local-Adaptive Face Recognition via Graph-based Meta-Clustering and Regularized Adaptation

Wenbin Zhu[1*]   Chien-Yi Wang[2*]   Kuan-Lun Tseng[2]   Shang-Hong Lai[2]   Baoyuan Wang[3]

[1]Microsoft Cloud and AI   [2]Microsoft AI R&D Center, Taiwan   [3]Xiaobing.AI

{wenzh, chiwa, kutseng, shlai}@microsoft.com   zjuwby@gmail.com

Abstract

Due to the rising concern of data privacy, it's reasonable to assume the local client data can't be transferred to a centralized server, nor their associated identity label is provided. To support continuous learning and fill the last-mile quality gap, we introduce a new problem setup called "local-adaptive face recognition (LaFR)". Leveraging the environment-specific local data after the deployment of the initial global model, LaFR aims at getting optimal performance by training local-adapted models automatically and un-supervisely, as opposed to fixing their initial global model. We achieve this by a newly proposed embedding cluster model based on Graph Convolution Network (GCN), which is trained via meta-optimization procedure. Compared with previous works, our meta-clustering model can generalize well in unseen local environments. With the pseudo identity labels from the clustering results, we further introduce novel regularization techniques to improve the model adaptation performance. Extensive experiments on racial and internal sensor adaptation demonstrate that our proposed solution is more effective for adapting face recognition models in each specific environment. Meanwhile, we show that LaFR can further improve the global model by a simple federated aggregation over the updated local models.

1. Introduction

Face recognition [30] has been commercialized widely for a variety of applications, such as FaceID, surveillance monitoring. The COVID-19 pandemic even accelerates the biometric technologies for touch-less solutions, such as face recognition enabled payment and access control. Although remarkable progress has been achieved lately, one has to admit that face recognition still hasn't been fully solved. Among many other remaining challenges (i.e., vulnerability for adversarial attack [4]), how to scale up the repre-
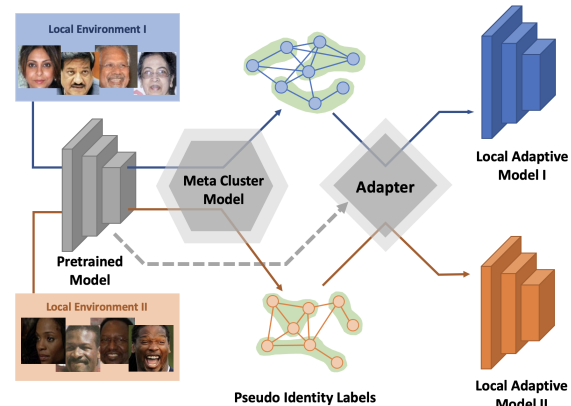

Figure 1. Local-Adaptive Face Recognition (LaFR): For each local environment, a specialized model is produced by the adapter module with only the pre-trained model and images from the environment. Note that there is no real identity label associated with the images, the meta-cluster model generates pseudo labels for robust model adaptation.

sentation learning to reduce the risk of fairness and bias to support various local environments becomes a more urgent challenge. As studied in previous works [14, 39, 44], such fairness and bias issues come from both algorithmic design and under-represented data distributions. For example, when the model is predominantly trained on RGB images, it generalizes poorly for images captured by Infrared cameras. Likewise, for a model pre-trained on Caucasian only, it performs substantially worse for African and Indian.

While it is worthwhile to push the domain-invariant face recognition [10, 25] with the hope of generalizing to everywhere without adaptation, it is arguably that the challenges for real-world scenarios could be more than we expected. So the question is: given an imperfect pre-trained model, how can we improve and fill the last-mile performance gap for each local environment and thereafter scale up the process? In this paper, we are interested in studying how to properly adapt the pre-trained model to a "specialized" one that tailors for the specific environment in an automatic and unsupervised manner. Here, the "environment" could be defined broadly, including a specific new

---

* indicates equal contribution.

camera sensor (i.e. an infrared camera with particular wavelength), a unique identity distribution with racial bias, or a physical environment that has unique camera placement and lighting condition, etc. We call such a problem setup as "Local-Adaptive Face Recognition (LaFR)", whose workflow (see Fig. 1) starts from an imperfect pre-trained global model deployed to a specific environment, where it accumulates some amount of new data. It then applies unsupervised adaptation technique to adapt the initial model locally, hence no data is transferred to server. Finally, after the adaptation, the new model is expected to perform much better than the initial global model as it is trained to tailor that environment. As an optional step, Federated Learning [24] is further employed to aggregate many such local models in a secure manner. Therefore, "LaFR" essentially provides a way to scale up the representation learning and model generalization via such "dual-loop" paradigm.

Although unsupervised domain adaptation (UDA) [37] has been widely studied in person re-identification (re-ID) [7, 18–20, 28, 43, 47], it is much less explored in face recognition except [33, 39]. Most of those works either designed special for person re-ID [7], or their setups require both source and target dataset to be available during the adaptation stage, or they only aim at closed-set problem. Moreover, person re-ID works heavily depend on variants of triplet loss, as we show in prior works, there are more robust losses (i.e., CircleLoss [29]) that proved to work better for face recognition.

To overcome the challenges, we first introduce a graph-based meta-clustering algorithm designed to predict pseudo labels for any unlabelled dataset. To do this, we collect a set of labeled datasets from multiple domains and extract their face embeddings from the given pre-trained model, we then apply Graph Convolution Network (GCN) to model the non-convex structure relationship for face embeddings within each set, which is trained efficiently through meta-learning [6] to make the cluster prediction more generalizable for the unseen dataset. Secondly, to better facilitate the transfer learning, we introduce a new technique by transferring the representation of class (pseudo label) center from the pre-trained model to the classifier of the new model and keep it fixed while only fine-tuning the feature representation in the context of margin-based training objectives such as [3,29,34]. Moreover, instead of regularizing the feature distance (commonly used in knowledge distillation [8,13,27]), we regularize the network weights to ensure a small deviation between the pre-trained and the new local model.

To summarize, we make the following major contributions: (1) we introduce a novel unsupervised model adaptation problem setup for face recognition, we argue that it's practical yet scalable motivated from both continuous learning and data privacy concerns; (2) we use graph convolution

network (GCN) to model the dataset structure and predict the clustering labels, through a meta learning framework; (3) Our novel regularized center transfer (RCT) technique can significantly reduce the risk of overfitting and improve transfer learning performance for even smaller datasets; (4) Experiments show that our entire solution not only outperform other strong baselines for local adaption but also enable the federated learning to further improve the global model.

2. Related Works

Generalized Face Recognition   Most of the recent works focus on generalizing the representation power through novel loss functions, such as CircleLoss [29], Arcface [3], Regularface [5] and AM-SoftMax [34], with both solid theoretical foundation and remarkable empirical results are reported in SOTA models. A set of other new works target to improve the performance from the data distribution perspective, including data augmentation [44] for underrepresented classes, improved training strategy for unbalanced and long-tailed data [45], and uncertainty modeling for noisy training set [2]. More recently, generalized representation learning has received more and more attention. [26] learns universal representations to tackle pose, resolution, and occlusion variation, while [10] uses meta-learning to make it easily generalized to the unseen domain without requiring adaptation. [35] proposes to use the adversarial decorrelation technique to make the identity representation invariant to age information. Causal relationship modeling through Invariant Risk Minimization [1] is a promising direction to address the out-of-distribution (OOD) challenge, yet, no impressive results have been reported so far for face recognition. Our method is both compatible and complementary with all those latest works. Improving the generalization capability of the pre-trained model will provide a better initialization for subsequent adaptation.

Unsupervised Domain Adaptation (UDA)   There is a large body of research works in unsupervised domain adaption [37] for person re-identification [7, 18–20, 28, 43, 47], semantic segmentation [31], image classifications [22, 33, 36] and attribute recognition [15]. However, most of those approaches either target for closed-set setting or require both source and target to be available during the adaptation to align the feature distribution through maximum mean discrepancy (MMD) [32]. Conceptually, a clustering technique is first needed to assign a pseudo label to each image before the training. Prior person re-ID works have explored various heuristic methods such as bottom-up cluster [19] and tracklet-based clustering [18]. In contrast, our clustering is achieved through a graph convolution network (GCN) to learn the intrinsic structure for each unseen data-set via meta-learning. To our best knowledge, there is little work
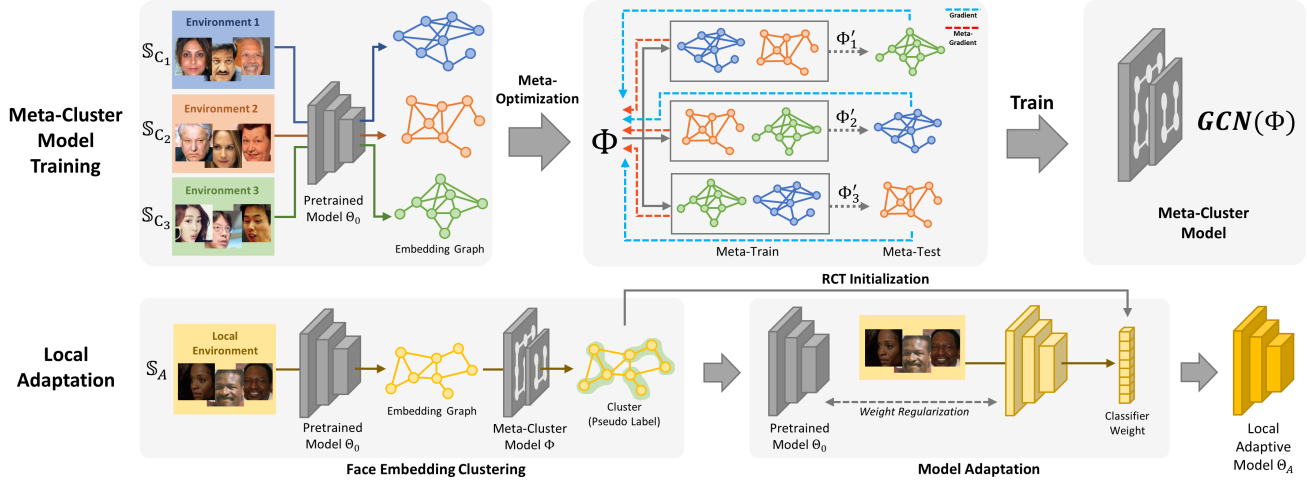
Figure 2. Overview of our Local-Adaptive face recognition (LaFR) framework. **Meta-Cluster Model Training:** A Graph Convolution Network (GCN) based embedding cluster model is trained with the meta-optimization procedure to generalize well in unseen local environments. **Test-time Local Adaptation:** In the local adaptation phase, we obtain the pseudo identity label of local face images from the meta-cluster model. Given the pre-trained face recognition model, local images, and their corresponding pseudo labels, the proposed Regularized Center transfer (RCT) technique can adapt the model more robustly to produce a specialized face recognition model.

of applying UDA for face recognition [23, 27]. Perhaps the most closely related work is [39] in which a traditional MMD loss is required to be employed. However, as we argued before, in the context of "LaFR", the source data is not available during the adaptation due to the data privacy and security concern.

Graph Convolution Network based Face Clustering Face clustering is another fundamental problem in computer vision, which is also quite related to our work. A recent series of studies [41, 42, 46] show that supervised clustering performs significantly better than the conventional unsupervised ones such as K-Means [21] and Spectral-Clustering [16]. The reason behind is that these Graph Convolution Network (GCN) [17] based methods are more capable of finding the local non-convex structures. Inspired by such remarkable progress, we move one step further and marry GCN-based clustering with meta-learning [6] to make the resulting clustering model more adaptive to the unseen dataset.

3. Overview

Given an imperfect pre-trained model Θ_0, our whole system requires a labeled dataset $\mathbb{S}_C$ in the global-end, which consists of a few subsets $(\mathbb{S}_{C_1}, ..., \mathbb{S}_{C_k})$ from multiple domains to train a novel graph-based meta-clustering model Φ. Once trained, given any unlabelled new dataset $\mathbb{S}_A$ on a specific local environment, we will feed it into Φ to obtain the pseudo label for each image. Then, we apply our proposed regularized center transfer (RCT) technique to perform the adaptation and produce a specialized local model

Θ_A tailors for that environment. Figure 2 illustrates the framework of our system.

4. Graph-based Meta-Clustering Learning

4.1. Preliminary of Graph Convolution Network

Given a face embedding set $\mathbb{S} = \{f_i, y_i\}_{i=1}^N$, which y_i is the corresponding identity label of each embedding. We first define a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ that connects each image with its K-NN neighbor, and the affinity matrix $\mathcal{A}$ is computed by cosine similarity between f_i and its neighbor $f_j (j \in \mathcal{N}_i)$, so $\mathcal{A}_{ij} = a_{ij} = \cos(f_i, f_j)$ and $\mathcal{N}_i$ represents the neighbors of f_i. Let us use $F_0 \in \mathbb{R}^{N \times d_{in}}$ to represent the input feature embedding for a Graph Convolution Network(GCN), then after l layers of convolutions on the graph, the feature embedding $F_l \in \mathbb{R}^{N \times d_{out}}$ becomes

$$F_l = \sigma(g(\bar{A}, F_{l-1})W_{l-1}^{\mathcal{G}}) \tag{1}$$

where $\bar{A} = \bar{D}^{-1}(A + I)$ and $\bar{D}_{ii} = \sum_j (A + I)_j$ is a diagonal degree matrix, $W_{l-1}^{\mathcal{G}} \in \mathbb{R}^{2 \times d_{in} \times d_{out}}$ is a trainable matrix to transfer the input embedding into a new space, σ is the nonlinear activation function(i.e., ReLU), and $g(\cdot, \cdot)$ is a concatenation function following previous works [41, 42, 46], which is defined as

$$g(\bar{A}, F_{l-1}) = [(F_{l-1})^T, (\bar{A}F_{l-1})^T]^T \tag{2}$$

In the context of this work, as we want to learn how to predict the face clustering for a given dataset, we let the last layer output a 1-D confidence value for each vertex. Assume c' represents the vector of confidence values for

each vertex, then

$$c' = F_l W^{\mathcal{G}} + b \qquad (3)$$

where $W^{\mathcal{G}}$ is the projection matrix and b is the bias, both are learnable. Again, following the design of "GCN-V" [41], we define the ground-truth confidence c_i as

$$c_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} (\mathbb{1}(y_j = y_i) - (\mathbb{1}(y_j \neq y_i))) \cdot a_{ij} \qquad (4)$$

where $\mathbb{1}()$ is an indicator function. Therefore, we minimize the following loss to train GCN:

$$\mathcal{L}^{\mathcal{G}}(\Phi) = \frac{1}{N} \sum_i^N |c_i - c_i'| \qquad (5)$$

where c_i' is the confidence value from c' for the corresponding vertex i, and Φ represents the parameters of the GCN model. Intuitively, a lower c_i means this vertex could lie near the boundary between two different clusters according to the local graph structure.

4.2. Meta-Clustering with GCN

By no means we are the first to apply GCN for face clustering given the impressive works done recently [41,42,46]. Nevertheless, our work is compatible yet builds on top of them by changing the training strategy into a meta-learning paradigm. We are interested in solving the aforementioned Local-Adaptive face recognition scenarios, where the pre-trained model is already deployed and their source data is not available anymore. Therefore, an automatic and unsupervised adaptation from a pre-trained model becomes a necessary and urgent problem, in order to massively scale-up the adaptation.

To train a meta-clustering model on GCN, we first collect a few labeled datasets $\mathbb{S}_{C_1}, ..., \mathbb{S}_{C_k}$ from C_K different domains and then perform domain-level sampling for outer-loop iteration, where we sample $C_K - 1$ domains data for the meta-train while the data from the remaining domain for the meta-test. Our goal is to let this GCN be more generalizable for any dataset from unseen domains. We describe the detailed training procedure in Algorithm 1. Specifically, We start with a randomly initialized GCN parameter setting Φ_0, and for each meta-train iteration, we perform the conventional GCN training via the following equation:

$$\Phi' = \Phi - \alpha \nabla_\Phi \mathcal{L}^{\mathcal{G}}_{mtr}(\Phi) \qquad (6)$$

where $\mathcal{L}^{\mathcal{G}}_{mtr}(\Phi)$ denotes the loss on the meta-train dataset. The model is then tested on the meta-test dataset. Similarly, we compute the corresponding loss as $\mathcal{L}^{\mathcal{G}}_{mte}(\Phi')$ with the updated model Φ'. The meta-model will then be updated jointly from both gradients (Line #8). This process is iterated until it reaches the maximum.

Algorithm 1 Graph-based Meta-Clustering
___
Input: Initialize Φ as Φ_0; Datasets $\mathbb{S}[\mathbb{S}_{C_1}, ..., \mathbb{S}_{C_k}]$; Hyper-parameters α, β, ξ
Output: Φ
1: **for** iter < MaxIter **do**
2:     **Split:** $\mathbb{S}$ into $\mathbb{S}_{mtr}$ and $\mathbb{S}_{mte}$ randomly
3:     **Meta-Train:**
4:     Train on $\mathbb{S}_{mtr}$: $\Phi' = \Phi - \alpha \nabla_\Phi \mathcal{L}^{\mathcal{G}}_{mtr}(\Phi)$
5:     **Meta-Test:**
6:     Compute loss on $\mathbb{S}_{mte}$: $\mathcal{L}^{\mathcal{G}}_{mte}(\Phi')$
7:     **Meta-optimization:** Update Φ
8:     $\Phi = \Phi - \beta(\nabla_\Phi \mathcal{L}^{\mathcal{G}}_{mtr}(\Phi) + \xi \nabla_\Phi \mathcal{L}^{\mathcal{G}}_{mte}(\Phi'))$
9: **end for**
___

5. Unsupervised Model Adaptation

As discussed before, our pipeline should only take the pre-trained model Θ_0 and an unlabeled dataset $\mathbb{S}_A$ ($\{x_i\}_{i=1}^N$) accumulated from the specific local environment as input. We will then run the above graph-based meta-clustering module to obtain the estimated number of pseudo "identity"(or class) labels as well as their belonging images. Let us denote (x_i, y_i) as one training image embedding with its associated pseudo ID: y_i, one can then employ any of the recent loss functions such as ArcFace [3], AM-Softmax [34] and CircleLoss [29] to train the adapted model Θ_A from Θ_0.

5.1. Preliminary of Face Recognition Training

The early idea behind face recognition is to treat it as a classification problem and apply standard softmax loss to train the deep face representation: $\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{W_{y_i}^T f_i + b_{y_i}}}{\sum_{j=1}^{\mathcal{C}} e^{W_j^T f_i + b_i}}$,where $f_i \in \mathbb{R}^d$ denotes the deep feature learned through the network for each image; $W_i \in \mathbb{R}^d$ represents the corresponding classifier vector from the last fully connected layer for label y_i, and b_i is the bias term. By normalizing both classifier vector $\| W_j \| = 1$ and $\| f_i \| = 1$, while adding a scale factor γ, we get $W_j^T f_i = \| w_j \| \| f_i \| \cos\theta_j = \cos\theta_j$, where θ_j is the angle between w_j and f_i, for simplicity, we also fix $b_j = 0$. To improve the generalization, recent study shows that adding different types of margins [29] [34] [3] can bring significant gain in many places. For example, ArcFace [3] adds angular margin while AM-softmax [34] adds cosine margin. Without loss of generality, we take AM-softmax to as the training objective: $\mathcal{L}(\Theta) = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\gamma(\cos(\theta_{y_i})-m)}}{e^{\gamma(\cos(\theta_{y_i})-m)}+\sum_{j\neq y_i}^{\mathcal{C}} e^{\gamma \cos\theta_j}}$, where $\mathcal{C}$ is the total number of classes, and m is the margin that needs empirically determined.

5.2. Regularized Center transfer (RCT)

We assume pre-trained representation (from Θ_0) already has strong discriminative power. After training, images belonging to the same identity tend to be clustered in a small region on the hyper-sphere, while still maintaining a cosine margin (defined as m) with images from different identities. If the pre-trained model was trained with a large number of different identities, the cluster area for each identity should have been squeezed on the hyper-sphere to reduce the intra-class variation. Let us denote $C_{y_i}^{\Theta_0}$ as the center of the face embedding for y_i on the pre-trained model, then $C_{y_i}^{\Theta_0} = \frac{1}{M_i} \sum_{k=1}^{M_i} \mathbb{1}(y_k = y_i) f_k^{\Theta_0}$, where M_i is the total number of images belonging to class y_i. Inspired by center loss [40], instead of learning f_i and its corresponding W_{y_i} from the new dataset from scratch, we propose to transfer this pre-trained class center as prior knowledge to the adapted model Θ_A, so basically we want to have $C_{y_i}^{\Theta_A}$ to be as close to $C_{y_i}^{\Theta_0}$ as possible during the adaptation. In the context of AM-softmax [34] or other latest training objectives such as CircleLoss [29], we notice that classifier vector W_{y_i} is also close to $C_{y_i}^{\Theta_0}$ even though they are not exactly the same. Therefore, in practice, to simplify the learning process, we directly use $C_{y_i}^{\Theta_0}$ to initialize each corresponding W_{y_i}, rather than learning it from scratch. To further reduce the overfitting risk when adapting to a small dataset, we add another model regularization term to let Θ_A not deviate too much from Θ_0. Our final loss function is defined as follows:

$$
\mathcal{L}(\Theta_A) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{\gamma(\cos(\theta_{y_i}) - m)}}{e^{\gamma(\cos(\theta_{y_i}) - m)} + \sum_{j \neq y_i}^{\mathcal{C}} e^{\gamma \cos \theta_j}}
$$
$$
+ \lambda \parallel \Theta_A - \Theta_0 \parallel_2^2
$$
(7)

subject to

$$
\begin{aligned}
W &= W^* / \parallel W^* \parallel, \\
f &= f^* / \parallel f^* \parallel, \\
\cos \theta_j &= W_j^T f_i, \\
W_{y_i} &= C_{y_i}^{\Theta_0}
\end{aligned}
$$
(8)

where W is the normalized classifier matrix, x is the normalized feature vector for each image, λ is a hyperparameter to trade-off between the loss on the new dataset and the model regularization term. During the model adaptation training, we initialize each W_{y_i} with $C_{y_i}^{\Theta_0}$ and keep it fixed and only fine-tune the feature representation. Compared with standard transfer learning without RCT, our regularized center transfer can better preserve the pre-trained representation, especially for small datasets, therefore reducing the risk of overfitting.

5.3. Federated Aggregation

As we argued in Sec. 1, with the rising concern of data privacy, it is reasonable to assume that training data would be highly de-centralized across different clients in the future, where model adaptation is conducted first locally at each client and later on being aggregated via federated learning [24] in a secure way. We argue that our proposed unsupervised modal adaptation is designed to further facilitate such fully automated "dual-loop" learning paradigm. To reduce the risk of losing privacy and adversarial attack, we remove the top classification layer and only transfer the backbone model parameters between each client and server and conduct simple model averaging, we denote such partial averaging setup as "FedPav" in our experiments [48].

6. Experimental Settings

6.1. Datasets and Protocols

In each Local-Adaptive face recognition (LaFR) protocol, the pre-trained face recognition model is trained with base dataset $\mathbb{S}_B = \{x_i^B, y_i^B\}$, and the meta-cluster GCN model is trained with multiple non-overlapped labeled datasets $\mathbb{S}_C = \{x_i^{C_1}, y_i^{C_1}\} \cup \{x_i^{C_2}, y_i^{C_2}\} \cup ...\{x_i^{C_k}, y_i^{C_k}\}$. While deploying the model in the specific scenario, the model is adapted with unlabeled dataset $\mathbb{S}_A = \{x_i^A\}$ from the environment. The final face recognition performance is evaluated on the testing dataset $\mathbb{S}_T = \{x_i^T, y_i^T\}$. Here, x_i and y_i represent the i-th face image and the corresponding identity label, respectively. Note that the base dataset $\mathbb{S}_B$ is not available during clustering and adaptation.

Adapt to Different Local Races The existence of face image distribution shift between different races has been proved in related works [39] [38], so we re-organized the racial datasets collected by [39] to build the LaFR protocols. We leverage the Caucasian dataset from BUPT-Transferface [39] as $\mathbb{S}_B$ for base face recognition model training, and racial faces in-the-wild (RFW) dataset [39] is used as $\mathbb{S}_T$ for testing after adaptation. BUPT-Balancedface [39], which has non-overlapped 7k subjects for each race, is leveraged for meta-cluster model training and further adaptation. The dataset details are summarized in Table 1. We build three leave-one-out protocols, which one of the races from {African, Asian, Indian} is selected as the local target, and dataset from other races can be used for meta-cluster GCN model training.

Adapt to Different Local Infrared Sensors Different wavelengths have different penetration rates into human skins, leading to different contrasts on the IR face images. Most cameras are sensitive to 850nm infrared light and generally considered to have higher contrast than 940nm infrared light. Meanwhile, different cameras also adapt different Image Signal Processors (ISP), so the way to handle

Dataset	Race	# Subjects	# Images
$\mathbb{S}_B$	Caucasian	10000	468139
$\mathbb{S}_C, \mathbb{S}_A$	Indian	7000	275095
	Asian	7000	325475
	African	7000	324376
$\mathbb{S}_T$	Indian	2984	10308
	Asian	2492	9688
	African	2995	10415

Table 1. Statistics of the datasets for races.

Dataset	Sensor	# Subjects	# Images
$\mathbb{S}_B$	RGB	94,430	5,179,510
$\mathbb{S}_C, \mathbb{S}_A$	IR-A	430	57,818
	IR-B	250	34,592
	IR-C	220	30,689
	IR-D	400	55,312
$\mathbb{S}_T$	IR-A	210	28,916
	IR-B	3,372	27,155
	IR-C	200	14,035
	IR-D	235	32,388

Table 2. Statistics of the datasets for sensors.

tone-mapping, noise reduction, and blurriness are different too, which causes another type of significant appearance distribution bias across different sensors. In this paper, we are interested in studying how to adapt a pre-trained RGB face recognition model to different local environments, each of which mimics a specific infrared camera enabled face recognition scenario. As there is no public infrared camera dataset with different ISP or wavelength, we collected an internal dataset with four different infrared camera sensors that capture infrared wavelength ranging from 850 nm to 940 nm. Figure 3 shows a few typical examples sampled from our four datasets, whose details are summarized in Table 2. We partition the collected dataset by identity to form $\mathbb{S}_C$, $\mathbb{S}_A$, and $\mathbb{S}_T$, and we leverage commonly used RGB dataset MS-1M [12] as $\mathbb{S}_B$. Similarly, we build four leave-one-out protocols, where we use datasets from 3 sensors to train our proposed meta-clustering while the remaining one for both adaptation and final testing.

6.2. Evaluation Metrics

Face Embedding Clustering Given the ground truth label from adaptation datasets $\mathbb{S}_A$, the intermediate face clustering result from the GCN model can be evaluated to indicate the accuracy of identity label assignment during adaptation. Two common clustering metrics [41, 42]: *Pairwise F-score* (F_P) and *B-Cubed F-score* (F_B) are used, which calculate the harmonic mean of precision and recall. The metric F_p puts relatively more emphasis on large face clusters, while F_B weights clusters linearly based on their size.

Local Adaptive Face Recognition The face recognition performance in the adaptation target is evaluated with standard face recognition metrics. In race adaptation protocols, we calculate the verification accuracy on 6000 difficult pairs
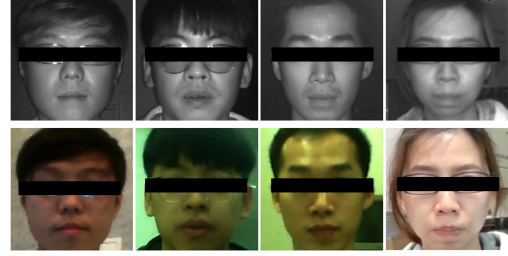

Figure 3. **Top:** Infrared Images captured by 4 different infrared cameras (the first two capture 940nm wavelength, the other two 850nm); **Bottom:** Their corresponding RGB images captured by another 4 RGB cameras.

selected by [39]. In sensor adaptation protocols, we report the False Non-Match Rates where the False Match Rate is 1e-6 (FNMR@FMR=1e-6).

6.3. Implementation Details

Meta-Clustering with GCN Training For GCN model training, we set the meta learning rate α to 0.1, the outer loop learning rate β to 0.1, and the meta loss weight ξ to 1.0 to conduct the meta optimization. In all experiments, we set the momentum to 0.9 and train 30000 iterations with the SGD optimizer. In each iteration, we randomly select one dataset from $\mathbb{S}_C$ as the meta-test set and the other datasets as the meta-train set. It takes around 7 hours to train on one TITAN X GPU.

Model Adaptation After clustering, we follow [41] to set a threshold τ to 0.8 to cut off the edges with small similarities, and obtain the pseudo identity labels by simply connecting each vertex to their nearest neighbors. During model adaptation with Regularized Center Transfer (RCT), we follow Equ. 7 to calculate the face embeddings center of each identity according to the pseudo labels. We set the model regularization λ to 0.1 and employ CircleLoss [29] as the classification loss. In all adaptation experiments, we set the learning rate to 0.001 and train for 50 epochs. It takes around 5 hours to train on one TITAN X GPU.

7. Experimental Results

7.1. Adapt to Different Races

Face Embedding Clustering With the ground truth identity label (not available during clustering and adaptation) from the dataset $\mathbb{S}_A$, we can evaluate the intermediate face clustering results from different methods using the F-score metrics described in Sec. 6.2. We compare the clustering performance with the original GCN [41] method, which uses all the datasets in $\mathbb{S}_C$ to build a large graph and learn the embedding confidence prediction through the supervised training scheme described in [41]. Besides, we also compare with another distance-based clustering method [39], which connects all the embedding pairs whose

Methods	African	Asian	Indian
Distance-based [39]	0.0086 / 0.7282	0.0035 / 0.6267	0.6891 / 0.7481
GCN [41]	0.6115 / 0.8132	0.3492 / 0.6432	0.8559 / **0.8725**
Meta-GCN	**0.8129** / **0.8535**	**0.3768** / **0.6876**	**0.8849** / 0.8551

Table 3. Comparison of face embedding clustering performance on three racial adaptation protocols. Two common clustering metrics: Pairwise / Bcubed F-score (F_P/F_B) pairs are reported.

cosine distance between them is less than a fixed threshold. We report the result with the threshold 0.3, which has the best performance across all sampled thresholds. The clustering performance of racial adaptation protocols is shown in Table 3, and we denote our proposed meta-clustering method as "meta-GCN". From the results, we can observe that the simple distance-based clustering method [39] has a very low Pairwise F-score (F_P) but high Bcubed F-score (F_B) in some adaptation protocols, which means that it cannot handle large embedding clusters. Our proposed "meta-GCN" outperforms the original GCN [41] in most of the benchmarks. It indicates that the proposed meta-learning scheme can learn more generalized GCN parameters and cluster the face embeddings better in unseen local environments. The clustering output with higher F-scores produces cleaner pseudo identity labels, which benefit more to the model adaptation process.

RCT Technique for Model Adaptation To evaluate the effectiveness of our proposed Regularized Center Transfer (RCT) technique, we leverage the ground truth (GT) identity labels from the adaptation dataset $\mathbb{S}_A$ to perform model adaptation. We compare the result with standard transfer learning, which is adapting the face recognition model without regularizing the classifier during the adaptation process. From Table 4, we have the following observations. First, the pre-trained model cannot perform well on three local targets without adaptation due to the race gap. Second, standard fine-tuning could possibly harm the face recognition model (ex. Indian protocol) due to the risk of overfitting on small dataset, which is illustrated in Figure 4) where it shows RCT is superior to standard fine-tuning across different size of identities. Third, our regularization techniques can effectively prevent the local model from overfitting and optimize the representation better than fine-tuning by a large margin in all protocols. Furthermore, we conduct ablation experiments to study the effect of different loss functions and the regularization term used in RCT. Results from Table 5 demonstrate that: 1) Our RCT can work with other margin-based classification loss like AM-softmax [34], but perform better with the SOTA CircleLoss [29]. 2) The weight regularization term (controlled by λ) does contribute a lot to the final adaptation performance.

End-to-end Unsupervised LaFR We further conduct end-to-end Local-Adaptive face recognition (LaFR) experiments, which combine face embedding clustering meth-

Methods	Caucasian	African	Asian	Indian
Pre-trained	0.9512	0.8537	0.8633	0.9047
$\mathbb{S}_A$-GT + Fine-tune	-	0.8758	0.8705	0.8528
$\mathbb{S}_A$-GT + RCT	-	**0.9227**	**0.9212**	**0.9323**

Table 4. Deployment model performance comparison while using the ground-truth (GT) labels from adaptation dataset ($\mathbb{S}_A$) on three racial protocols. Verification accuracy on 6000 pairs of the testing dataset ($\mathbb{S}_T$) is reported.

Methods	African	Asian	Indian
Pre-trained	0.8537	0.8633	0.9047
RCT (AM-Softmax, λ=0.0)	0.9125	0.9050	0.9048
RCT (CircleLoss, λ=0.0)	0.9145	0.9033	0.9120
RCT (AM-Softmax, λ=0.1)	0.9158	0.9153	0.9278
RCT (CircleLoss, λ=0.1)	**0.9227**	**0.9212**	**0.9323**

Table 5. The influence of loss function choices and weight regularization λ on racial adaptation protocols. All experiments are conducted with "$\mathbb{S}_A$-GT + RCT" setting.

Methods	African	Asian	Indian
Pre-trained	0.8537	0.8633	0.9047
MFR [11]	0.8882	0.8768	0.9118
Distance-based [39] + RCT	0.7613	0.7488	0.8862
GCN [41] + RCT	0.8692	0.8717	0.8940
Meta-GCN + RCT	**0.8912**	**0.8773**	**0.9258**
$\mathbb{S}_A$-GT + RCT	0.9227	0.9212	0.9323

Table 6. Deployment model performance comparison while using pseudo identity labels from **different clustering methods** combined with RCT on three racial protocols. Verification accuracy on 6000 pairs of the testing dataset ($\mathbb{S}_T$) is reported.

ods (from Table 3) with the proposed RCT adaptation technique. The end-to-end LaFR performance on three race protocols are shown in Table 6. The result in the last row which adopts ground truth (GT) labels from $\mathbb{S}_A$ serves as the upper bound of end-to-end unsupervised LaFR methods. We also re-implement and compare with the recent generalized face recognition method MFR [11] in our protocol. The dataset in $\mathbb{S}_C$ are leveraged into the meta-optimization process to obtain a better universal model. Although MFR works better than standard pretraining, we still observe the performance gap between MFR and our adaptation results, indicating the value of local adaptation for optimized performance. Meanwhile, Table 6 also shows that our proposed "meta-GCN" outperforms both distance-based [39] and regular GCN [41] based clustering methods in our protocols, which proves the effectiveness of the graph representation as well as the meta learning on top of the GCN.

7.2. Adapt to Different Sensors

We conduct similar end-to-end LaFR experiments on four infrared sensor adaptation protocols, note that those sensors are from real industry face recognition systems. From Table 7 and Figure 5, we have the following observations. First, as the upper bound for tested end-to-end LaFR methods, "$\mathbb{S}_A$-GT + RCT" can reduce the adaptation non-match error up to 70% compared with the strong RGB pre-
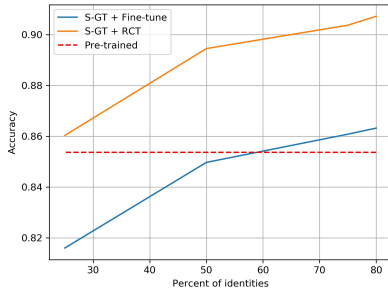
Figure 4. Performance comparison between $\mathbb{S}_A$-**GT + Fine-tune** and $\mathbb{S}_A$-**GT + RCT** under different percent of identities on the "African" adaptation protocol.

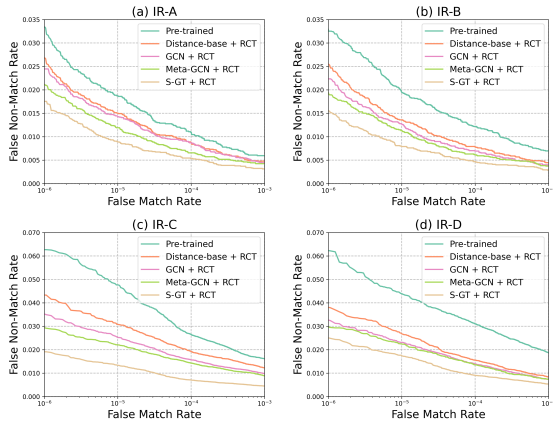Figure 5. The ROC curves of deployed models evaluated on all pairs of (a) IR-A, (b) IR-B, (c) IR-C, (d) IR-D protocols.

Label	Loss	FedPav	IR-A	IR-B	IR-C	IR-D
$\mathbb{S}_B$-GT	(Pre-trained)		0.0344	0.0330	0.0633	0.0626
SpC [9]	UCL [9]		0.0332	0.0298	0.0579	0.0623
SpC [9]	RCT		0.0256	0.0256	0.0443	0.0425
D-based [39]	RCT		0.0267	0.0254	0.0433	0.0384
GCN [41]	RCT		0.0245	0.0223	0.0358	0.0331
Meta-GCN	RCT		0.0212	0.0198	0.0298	0.0299
Meta-GCN	RCT	✓	**0.0207**	**0.0188**	**0.0274**	**0.0267**
$\mathbb{S}_A$-GT	RCT		0.0175	0.0158	0.0192	0.0251
$\mathbb{S}_A$-GT	RCT	✓	0.0175	0.0164	0.0187	0.0228

Table 7. Comparison between different LaFR methods on sensor adaptation protocols. The performance is measured by FNMR@FMR=1e-6, the lower the better.

trained model, which demonstrate the effectiveness of our adaptation regularization techniques for face recognition. Second, with limited number of identities collected in the environment (ex. 220 in IR-C), standard fine-tuning ("$\mathbb{S}_A$-GT + fine-tune") is prone to overfitting and cannot achieve better performance while sensor adaptation compared with the pre-trained model. Third, our proposed "Meta-GCN + RCT" achieves the best performance in all protocols among different clustering methods, which shows the generalization ability and robustness of our meta-clustering GCN model in different scenarios.

Algorithm 2 Model Adaptation in Federated Learning

Input: Pretrained model on server Θ_s^0; Client models $\Theta_{c_i}^0$; Client datasets D_i (i=1,2,...,K);
Output: Client models $\Theta_{c_i}^T$;
1: **for** each step t = 0 to T-1 **do**
2:     **for** each model i = 1 to K **do**
3:         $\Theta_{c_i}^{t+1} \leftarrow RCT(\Theta_s^t, \Theta_{c_i}^t, D_i)$
4:     **end for**
5:     $\Theta_s^{t+1} \leftarrow \frac{1}{K}\sum_{i=1}^K \Theta_{c_i}^t$
6: **end for**

Compared with SOTA Domain-adaptive ReID To show the necessity of our proposed solution, we further compare with the SOTA object Re-ID method [9]. We applied both their Self-paced Clustering(SpC) and the Unified Contrastive Loss(UCL) in our system. Even though SpC performs impressively on person reID task, their techniques are still inferior to our proposed Meta-GCN with RCT on the task of sensor adaptation for face recognition, which is likely caused by the unique challenging nature of face recognition task.

Experiments on Federated Learning We conduct experiments to verify the effectiveness of our LaFR in federated learning setting. Assuming there are K clients ($i = 1, 2, ..., K$), each associates with unlabeled private dataset D_i, and the face recognition model Θ_i^t at the optimization dual-loop step t. The continual model adaptation process within federated learning setup is shown in Algorithm 2. The training procedure contains 20 rounds of dual-loop, and each dual-loop consists of a local adaptation and a partial federated aggregation (FedPav) which averages the backbone parameters from the local clients. The results shown in Tab. 7 verify the effectiveness of the federated learning pipeline to iteratively improve the local face recognition models without breaching user's privacy on each client.

8. Conclusions

We introduce a new problem setup called "Local-Adaptive Face Recognition (LaFR)", which aims to produce specialized face recognition models tailored for each local environment. Our proposed graph-based meta-clustering model can better cluster the face embeddings for unseen environments, which provides cleaner identity labels in an unsupervised manner during adaptation. Combined with the proposed RCT module, our framework can robustly produce LaFR models adapted from imperfect pre-trained face recognition models. The effectiveness of our framework is proven on various protocols, including racial and sensor adaptation, with or without federated aggregation. We hope these efforts can open up future directions towards specialized face recognition models at scale.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2

[2] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *CVPR*, 2020. 2

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2, 4

[4] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *CVPR*, 2019. 1

[5] Yueqi Duan, Jiwen Lu, and Jie Zhou. Uniformface: Learning deep equidistributed representation for face recognition. In *CVPR*, 2019. 2

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 2, 3

[7] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, U. Uiuc, and T. Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6111–6120, 2019. 2

[8] S. Ge, S. Zhao, C. Li, and J. Li. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing*, 28(4):2051–2062, 2019. 2

[9] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020. 8

[10] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z. Li. Learning meta face recognition in unseen domains. In *CVPR*, 2020. 1, 2

[11] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *CVPR*, 2020. 7

[12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 6

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 2

[14] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *CoRR*, abs/1709.01450, 2017. 1

[15] Wen Ji, Kelei He, Jing Huo, Zheng Gu, and Yang Gao. Unsupervised domain attention adaptation network for caricature attribute recognition, 2020. 2

[16] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000. 3

[17] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017. 3

[18] Devinder Kumar, Parthipan Siva, Paul Marchwica, and Alexander Wong. Unsupervised domain adaptation in person re-id via k-reciprocal clustering and large-scale heterogeneous environment synthesis, 2020. 2

[19] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8738–8745, Jul. 2019. 2

[20] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang. Unsupervised person re-identification via cross-camera similarity exploration. *IEEE Transactions on Image Processing*, 29:5481–5490, 2020. 2

[21] Stuart Lloyd. Least squares quantization in pcm. In *IEEE transactions on information theory*, pages 28(2):129–137, 1982. 3

[22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[23] Z. Luo, J. Hu, W. Deng, and H. Shen. Deep unsupervised domain adaptation for face recognition. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2018. 3

[24] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016. 2, 5

[25] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K. Jain. Towards universal representation learning for deep face recognition. In *CVPR*, 2020. 1

[26] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K Jain. Towards universal representation learning for deep face recognition. In *CVPR*, 2020. 2

[27] Kihyuk Sohn, Sifei Liu, Guangyu Zhong, Xiang Yu, Ming-Hsuan Yang, and Manmohan Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. In *ICCV*, 2017. 2, 3

[28] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice, 2018. 2

[29] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, 2020. 2, 4, 5, 6, 7

[30] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1

[31] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez. Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10):7178–7193, 2020. 2

[32] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. In *Advances in Neural Information Processing Systems*, 2016. 2

[33] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2

[34] Feng Wang, Weiyang Liu, Haijun Liu, and Jian Cheng. Additive margin softmax for face verification. *arXiv preprint arXiv:1801.05599*, 2018. 2, 4, 5, 7

[35] Hao Wang, Dihong Gong, Zhifeng Li, and Wei Liu. Decorrelated adversarial learning for age-invariant face recognition. In *CVPR*, 2019. 2

[36] Hui Wang, Jian Tian, Songyuan Li, Hanbin Zhao, Qi Tian, Fei Wu, and Xi Li. Unsupervised domain adaptation for image classification via structure-conditioned adversarial learning, 2021. 2

[37] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *arXiv preprint arXiv:1802.03601*, 2018. 2

[38] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *CVPR*, 2020. 5

[39] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *ICCV*, 2019. 1, 2, 3, 5, 6, 7, 8

[40] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 5

[41] Lei Yang, Dapeng Chen, Xiaohang Zhan, Rui Zhao, Chen Change Loy, and Dahua Lin. Learning to cluster faces via confidence and connectivity estimation. In *CVPR*, 2020. 3, 4, 6, 7, 8

[42] Lei Yang, Xiaohang Zhan, Dapeng Chen, Junjie Yan, Chen Change Loy, and Dahua Lin. Learning to cluster faces on an affinity graph. In *CVPR*, 2019. 3, 4, 6

[43] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook, 2021. 2

[44] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, 2019. 1, 2

[45] Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Unequal-training for deep face recognition with long-tailed noisy data. In *CVPR*, 2019. 2

[46] Yali Li Zhongdao Wang, Liang Zheng and Shengjin Wang. Linkage-based face clustering via graph convolution network. In *CVPR*, 2019. 3, 4

[47] Weiming Zhuang, Yonggang Wen, Xuesen Zhang, Xin Gan, Daiying Yin, Dongzhan Zhou, Shuai Zhang, and Shuai Yi. Performance optimization of federated person re-identification via benchmark analysis. *Proceedings of the 28th ACM International Conference on Multimedia*, Oct 2020. 2

[48] Weiming Zhuang, Yonggang Wen, Xuesen Zhang, Xin Gan, Daiying Yin, Dongzhan Zhou, Shuai Zhang, and Shuai Yi. Performance optimization of federated person re-identification via benchmark analysis. *Proceedings of the 28th ACM International Conference on Multimedia*, Oct 2020. 5