Fast Adapting Without Forgetting for Face Recognition

Hao Liu, Xiangyu Zhu, *Member, IEEE*, Zhen Lei, *Senior Member, IEEE*,
Dong Cao, and Stan Z. Li, *Fellow, IEEE*

汇报人姓名：王岩

日期：2022.1.21

研究思路

The problem is:

- a well trained model on the data domain can only achieve good performance in this domain, it cannot cover new domains.

The common ways are as follows:

- Fine-tune the base model with the target-domain data to cope with new domain,which can cause catastrophic forgetting.

- Fine-tunes the model on both source and target domains simultaneously,it will take huge training time and data storage.

研究进展——已取得结果

- Single Exemplar Domain Incremental Learning(**SE-DIL**): a new task for a practical application of face recognition, which aims to quickly adapt the base model from source domain to the target domain and keep the performance on source domains.

- Fast Adapating without Forgetting(**FAwF**): a method to solve **SE-DIL** with three components: margin-based exemplar selection, prototype-based class extension and hard&soft knowledge distillation.

- **KidsFace**:a large-scale database of children faces with 12,444 identities,which is the first large-scale children database.

研究进展——

Fig. 1. The process of Single Exemplar Domain Incremental Learning. Starting with a well-trained base model, each time we encounter a new domain, it can adapt to the new domain and preserve the performance of the source domain, and finally, get superior generalization capabilities.

技术路线

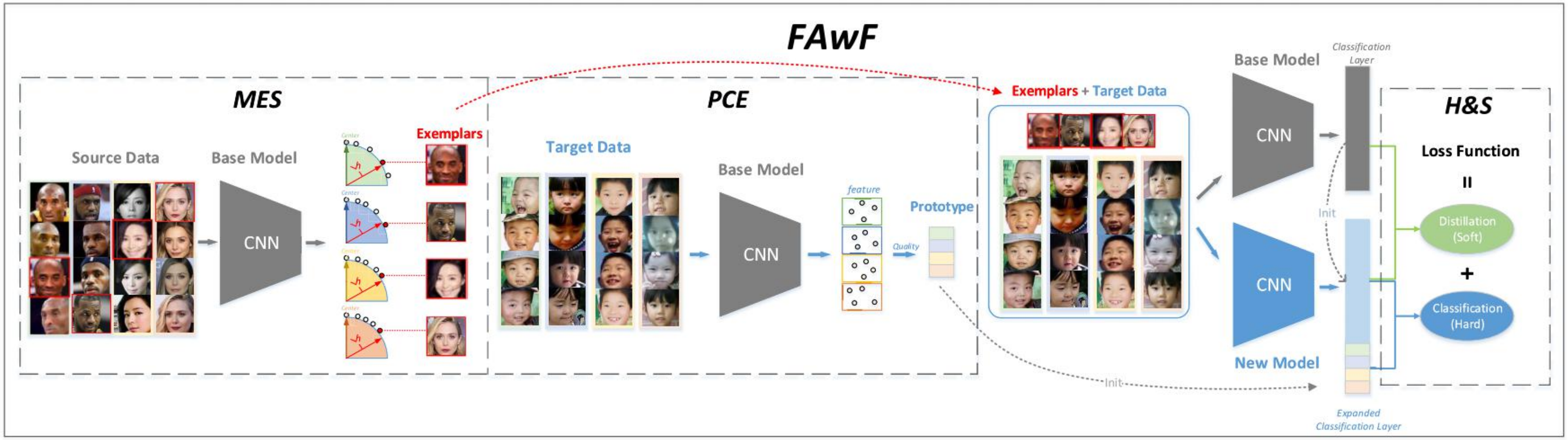
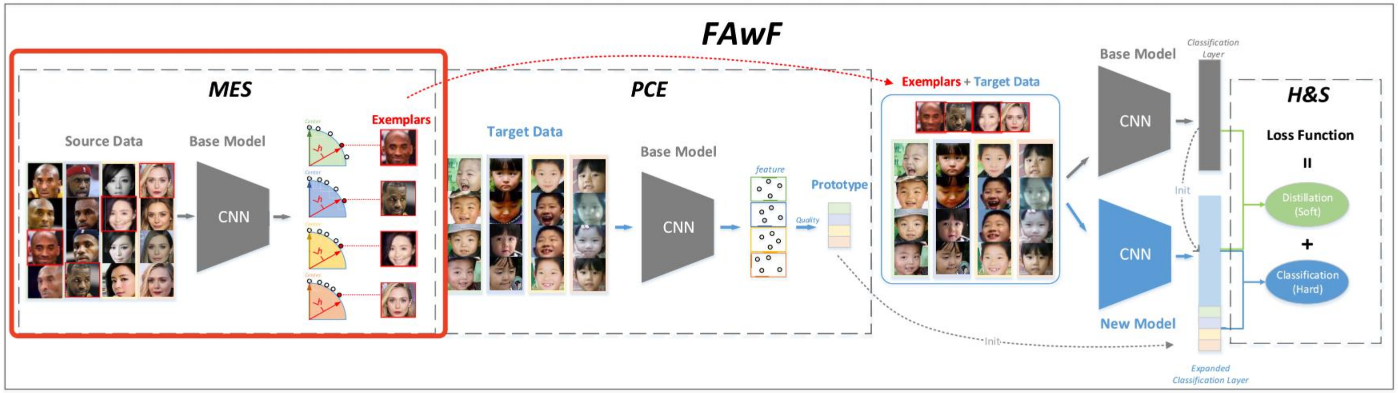


Fig. 3. Overview of our Fast Adapting without Forgetting. It consists of Margin-based Exemplar Selection (MES), Prototype-based Class Extension (PCE) and Hard&Soft Knowledge Distillation (H&S). The base model is not updated during training.

FAwF consists of three components:

- Margin-based Exemplar Selection(**MES**)
- Prototype-based Class Extension(**PCE**)
- Hard&Soft Knowledge Distillation(**H&S**)

技术路线



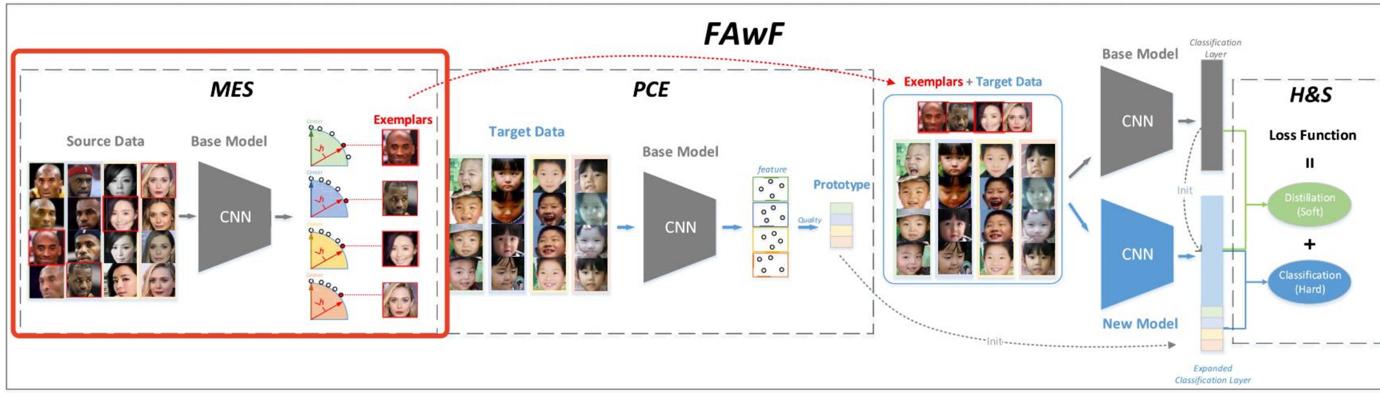
MES: select the most valuable exemplar one class to preserve source-domain knowledge.

The common way is to keep the samples that are as close to the class center.

MES only selects one sample a class to provide more diverse source domain intra-class information in target domain training to preserve source domain performance.

With a given margin h, select the sample whose distance from the class center is closest to h, this sample as the exemplar of this class.

技术路线



Algorithm 1 Margin-Based Exemplar Selection

Input : $Net(\theta_s)$

$D_s = \{(x_s^i, y_s^i), 1 \leq i \leq M_s, 1 \leq y_s^i \leq N_s\}$

Margin h

Output: Selected exemplars E_s

for *each sample* x_s^i *in* D_s **do**

    Extract the feature f_s^i of x_s^i from $Net(\theta_s)$

end

for $j = 1 \dots N_s$ **do**

    $c_j = Average(f_s^i), y_s^i = j$

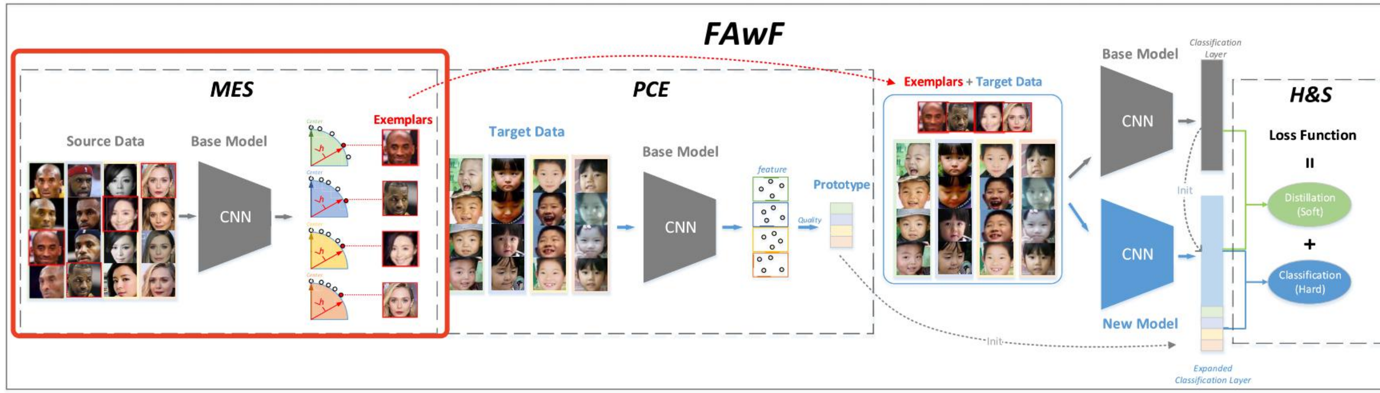
    $distance_s^i = ||f_s^i - c_j||, y_s^i = j$

    $e_j = x_s^{\arg\min(|distance_i - h|)}, y_s^i = j$

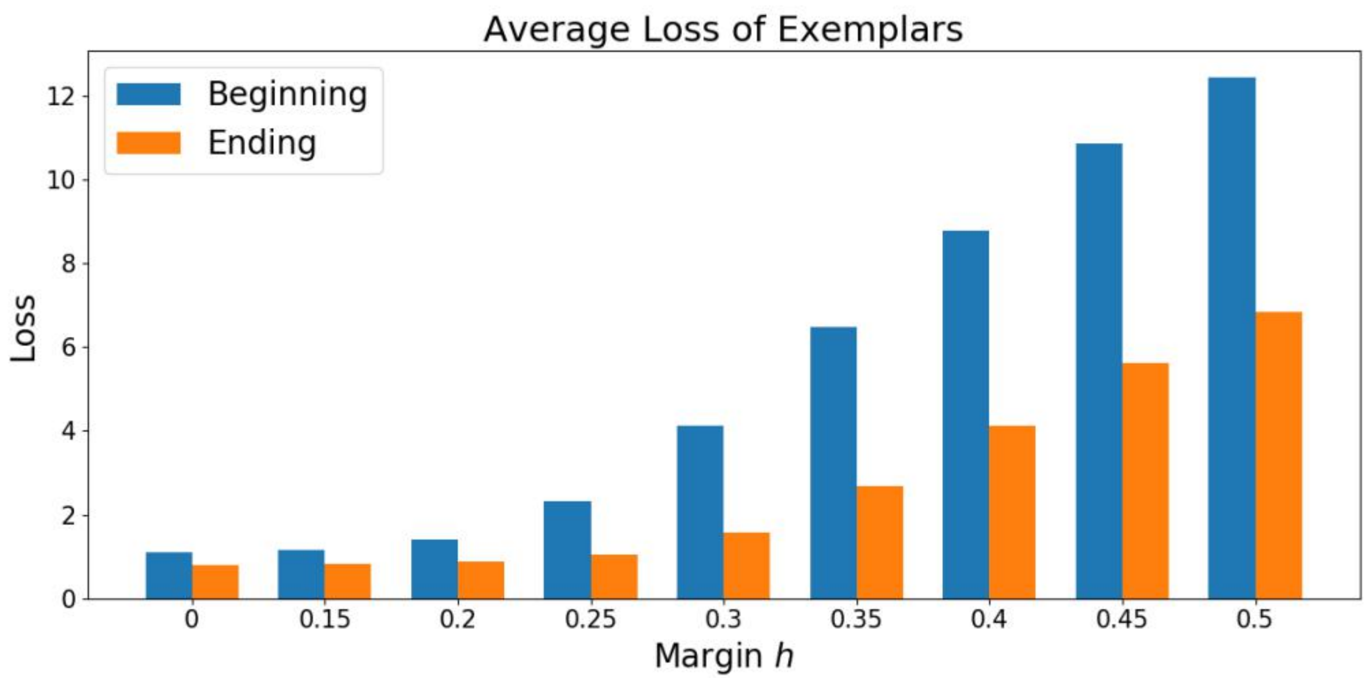
end

$E_s = \{e_j, 1 \leq j \leq N_s\}$

技术路线

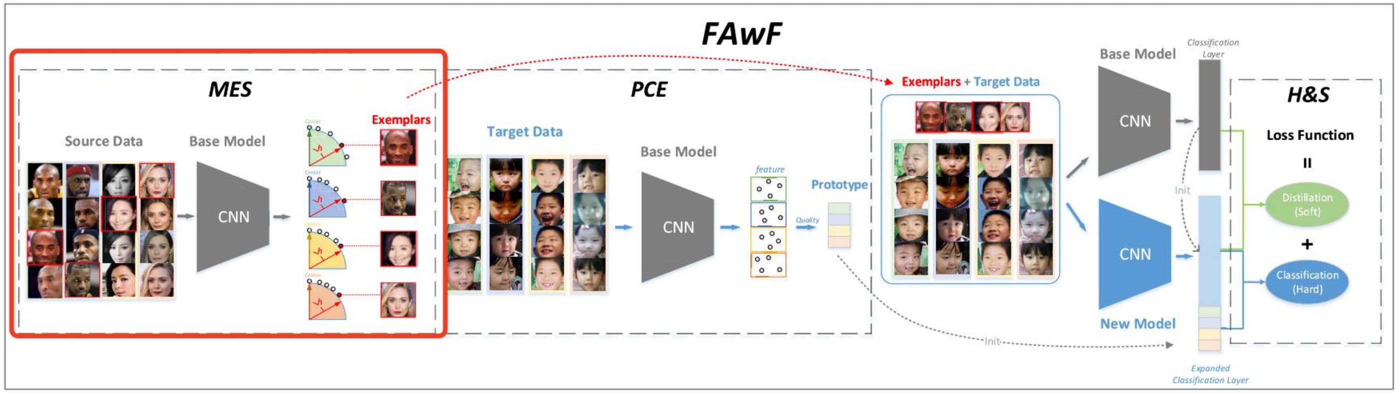


FAwF
MES
Source Data
Base Model
CNN
Exemplars
PCE
Target Data
Base Model
CNN
feature
Prototype
Exemplars + Target Data
Base Model
CNN
Classification Layer
Init
CNN
New Model
Init
Expanded Classification Layer
H&S
Loss Function
=
Distillation (Soft)
+
Classification (Hard)

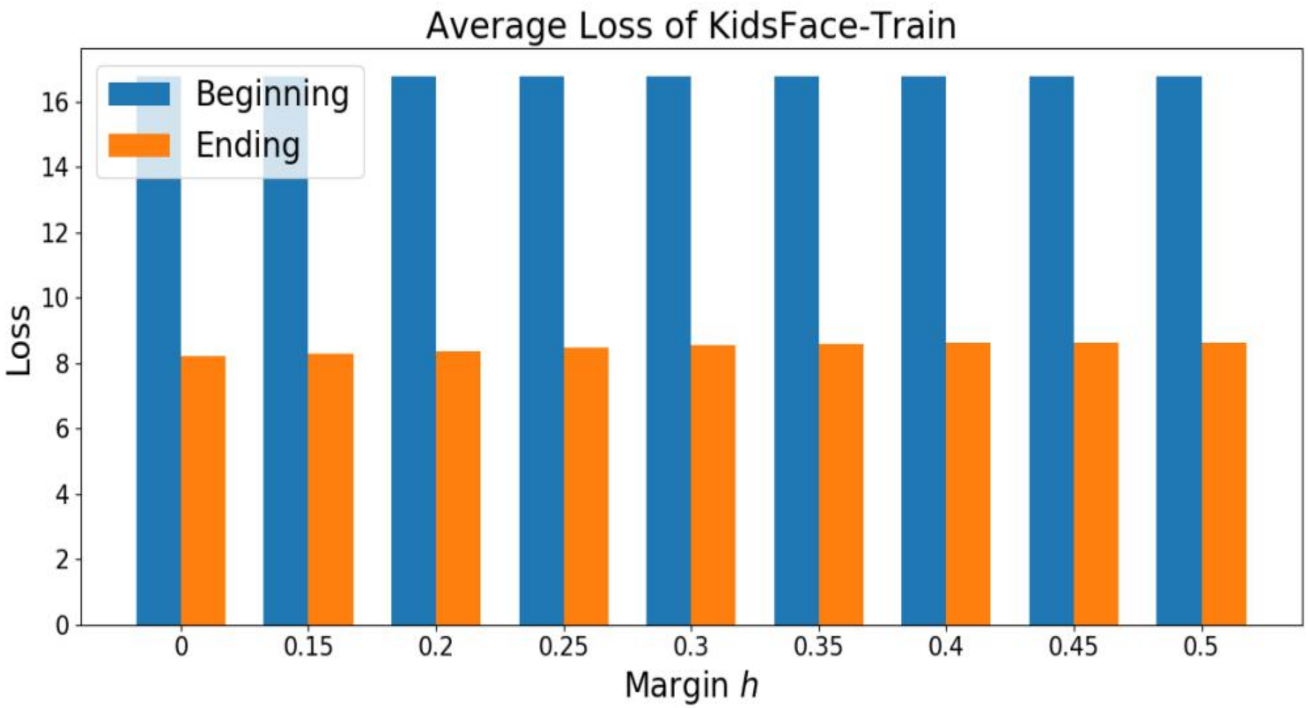


Average Loss of Exemplars
Beginning
Ending
Loss
0
2
4
6
8
10
12
0
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Margin h

技术路线

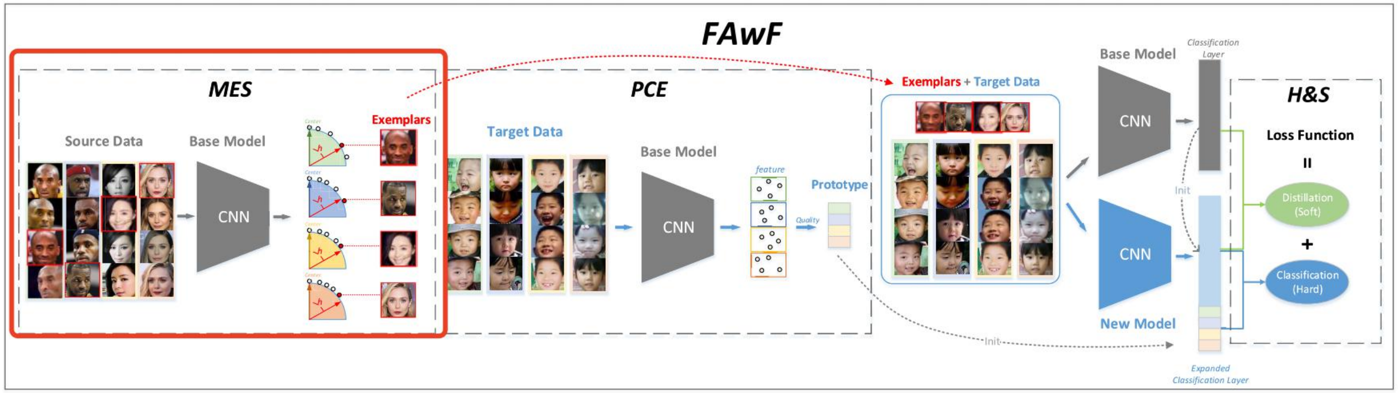


FAwF
MES
Source Data
Base Model
CNN
Exemplars
PCE
Target Data
Base Model
CNN
feature
Prototype
Exemplars + Target Data
Base Model
CNN
Classification Layer
New Model
CNN
Init
Expanded Classification Layer
H&S
Loss Function
Distillation (Soft)
+
Classification (Hard)

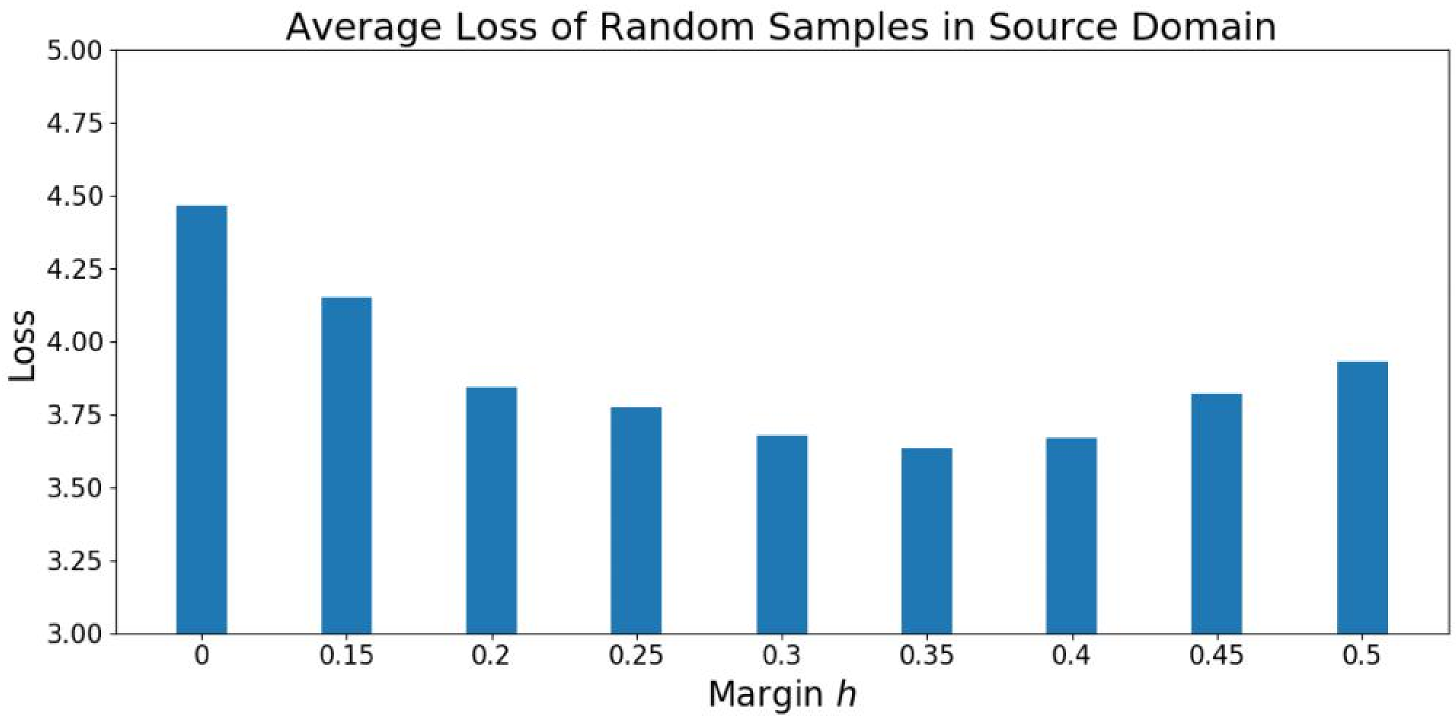


Average Loss of KidsFace-Train
Beginning
Ending
Loss
Margin h
0
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5

技术路线

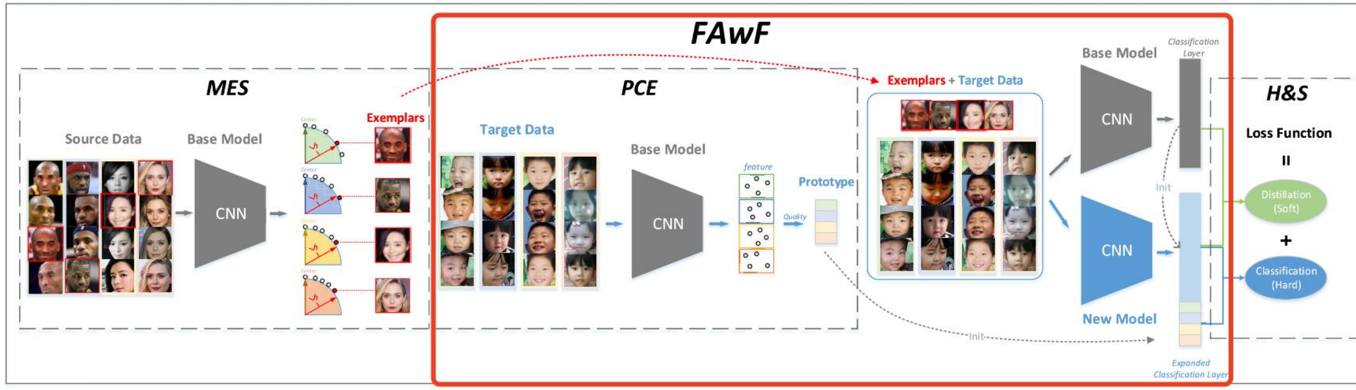


FAwF
MES
PCE
H&S
Source Data
Base Model
CNN
Exemplars
Target Data
Base Model
CNN
Prototype
Exemplars + Target Data
Base Model
CNN
New Model
CNN
Loss Function
Distillation (Soft)
+
Classification (Hard)

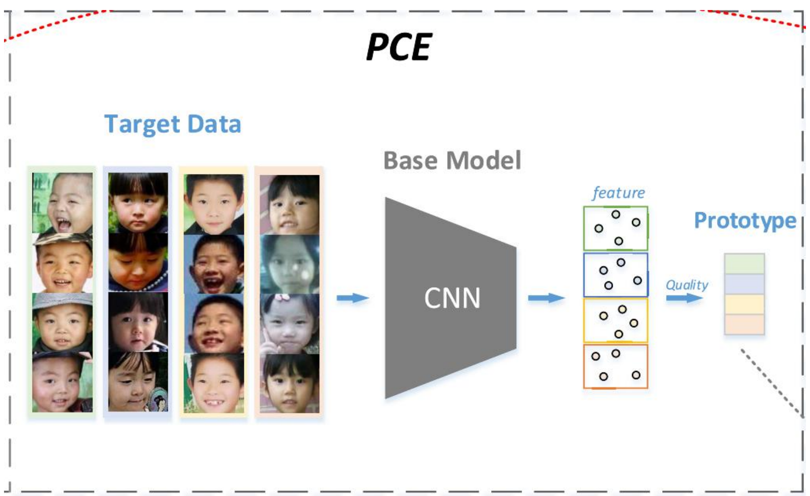


Average Loss of Random Samples in Source Domain
Loss
5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
0
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Margin h

技术路线



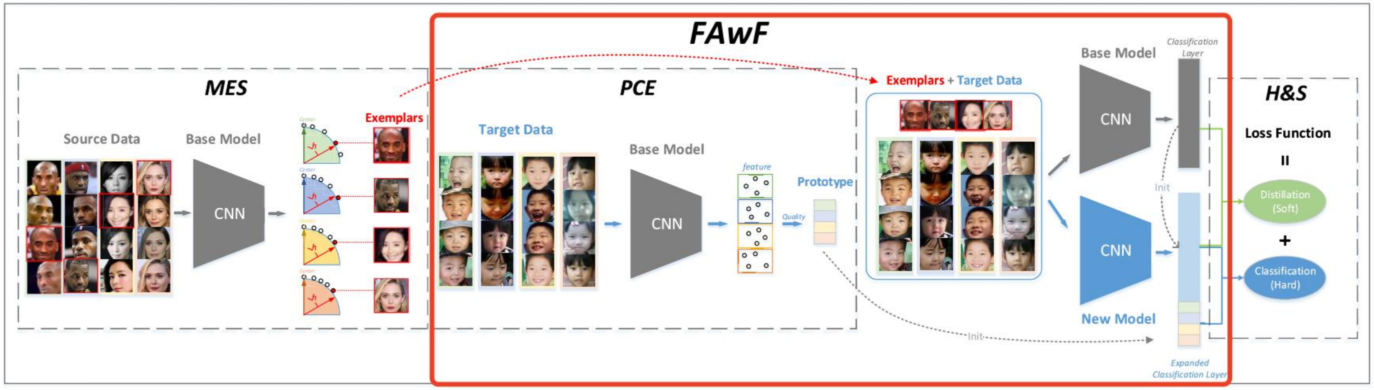
PCE： The classification layer needed to expanded when continuously adapted to new classes of new domains.



The incremental learning methods use random initialization for the weights of new classes.

take the class prototype to initialize the weights of new classes for the new domain in the classification layer.

技术路线



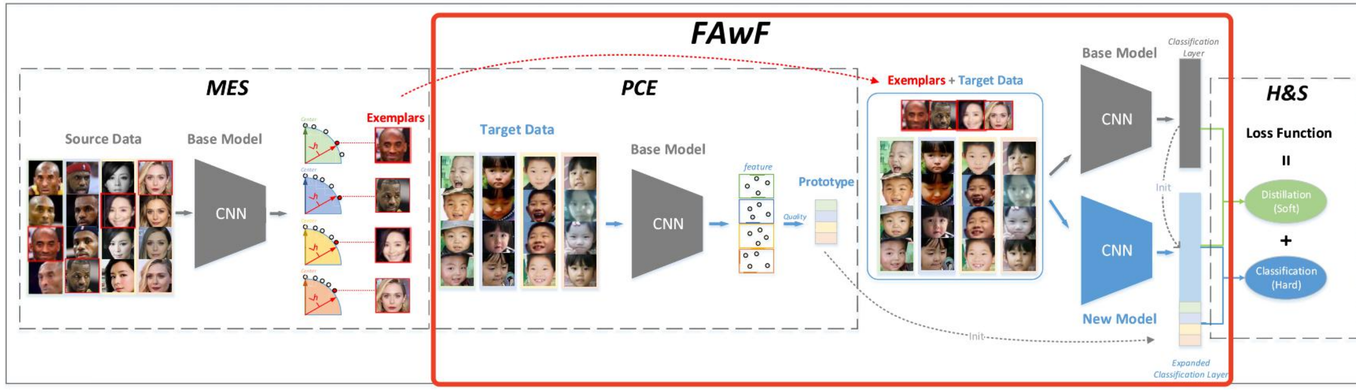
PCE: The classification layer needed to expanded when continuously adapted to new classes of new domains.

$$p_j = \frac{\sum_{y_t^i=j} f_t^i}{N_t}, \quad 1 \le j \le N_t \quad \longrightarrow \quad W_t = [p_1, p_2, \ldots, p_{N_t}]$$

extract features of all samples of the target domain, calculate the prototype to represent each new class.

$$p_j = \frac{\sum_{y_t^i=j} \mu_t^i f_t^i}{\sum_{y_t^i=j} \mu_t^i}, \quad 1 \le j \le N_t \qquad \mu_t^i = \| f_t^i \|$$

技术路线



Algorithm 2 Prototype-Based Class Extension

Input : $Net(\theta_s)$

$D_t = \{(x_t^i, y_t^i), 1 \le i \le M_t, 1 \le y_t^i \le N_t\}$

Output: Weight vectors W_t of target domain classes

for *each sample* x_t^i *in* D_t **do**

  Extract the feature f_t^i of x_t^i from $Net(\theta_s)$

end

for $j = 1 \ldots N_t$ **do**

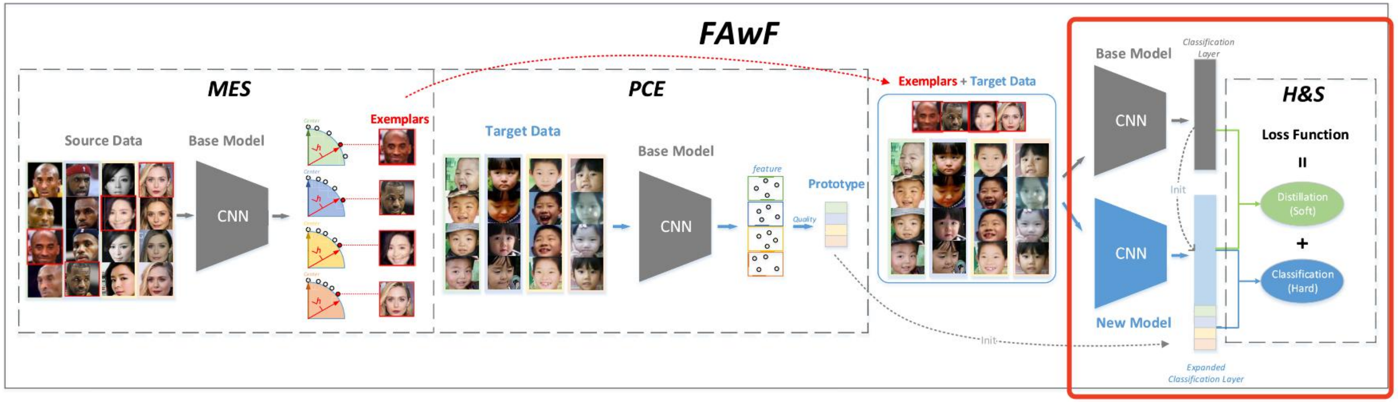
  $p_j = \frac{\sum_{y_t^i = j} \mu_t^i f_t^i}{\sum_{y_t^i = j} \mu_t^i}$, μ_t^i is quality factor

  $w_j = p_j$

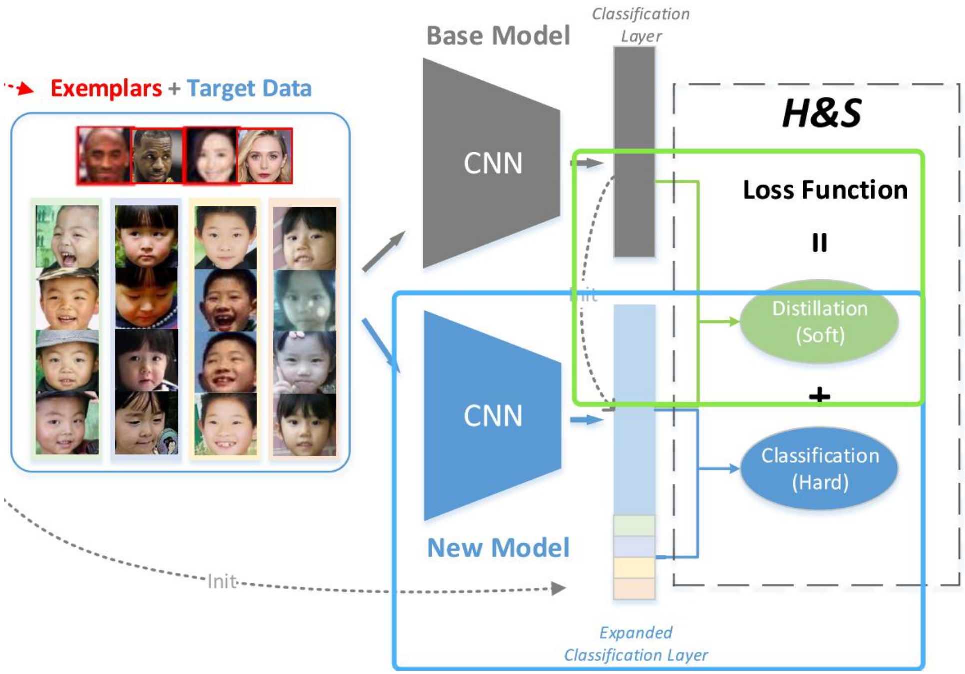
end

$W_t = \{w_j, 1 \le j \le N_t\}$

技术路线



Hard&Soft Knowledge Distillation:

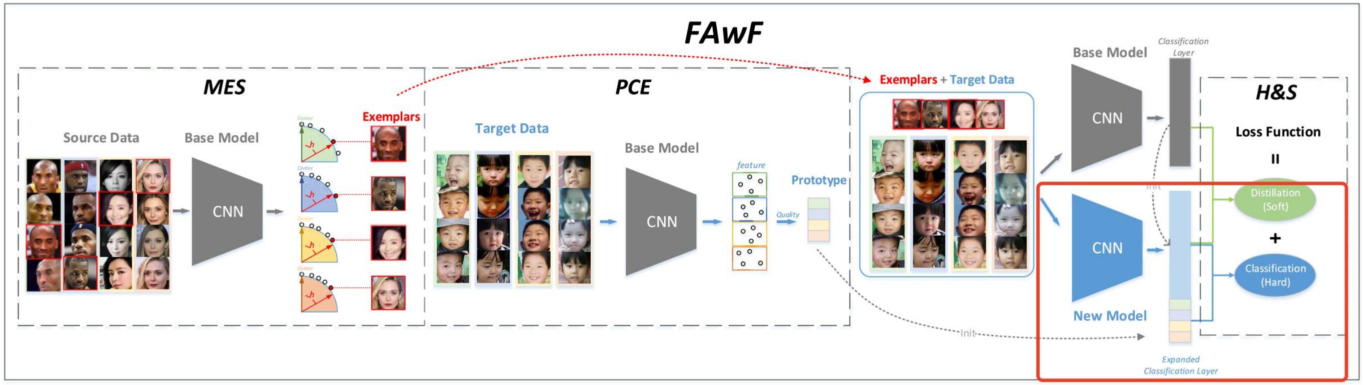


Loss combines the classification loss and the distilling loss.

$$L = L_c + \lambda \cdot L_d$$

The hard classification loss classifies exemplars in D_s and samples in D_t to their right labels.

技术路线



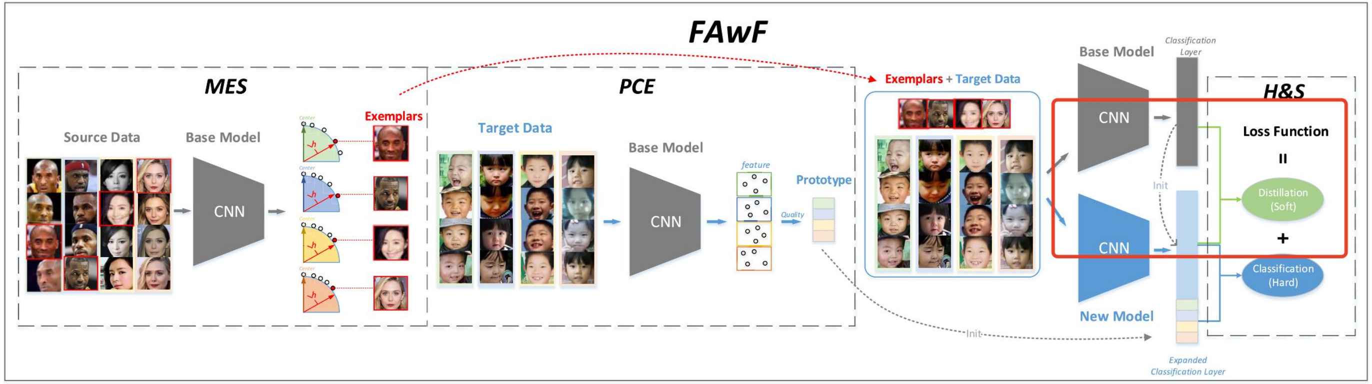
Use the **CosFace** for hard classification.

$$L_c = -\frac{1}{M_t + N_s} \times \sum_{i=1}^{M_t+N_s} \log \frac{e^{s(\cos(\theta_{iy^i})-m)}}{e^{s(cos(\theta_{iy^i})-m)} + \sum_{j=1, j\neq y^i}^{N_t} e^{s\cos(\theta_{ij})}} \tag{4}$$

$$\cos\theta_{ij} = w_j^{\mathrm{T}} f^i, \quad \|w_j\| = 1, \ \|f^i\| = 1 \tag{5}$$

where M_t is the number of samples of target domain, N_t is the number of target domain classes, w_j denotes the weight vector of class j, s is the scale factor and m is the margin parameter in CosFace.

技术路线



Take soft activation distilling loss to better retain the source domain information, the base model is used to guide the training.

Specifically, we denote the output logits of the base model and the new model as $\hat{\mathbf{o}}^{N_s|}(x) = [\hat{o}_1(x), \ldots, \hat{o}_{N_s}(x)]$ and $\mathbf{o}^{N_s+N_t}(x) = [o_1(x), \ldots, o_{N_s}(x), o_{N_s+1}(x), \ldots, o_{N_s+N_t}(x)]$,

$$L_d = -\frac{1}{M_t + N_s} \sum_{i=1}^{M_t+N_s} \sum_{j=1}^{N_s} \hat{\pi}_j(x_i) \log[\pi_j(x_i)] \tag{6}$$

$$\hat{\pi}_j(x_i) = \frac{e^{\hat{o}_j(x_i)/T}}{\sum_{k=1}^{N_s} e^{\hat{o}_k(x_i)/T}}, \quad \pi_j(x_i) = \frac{e^{o_j(x_i)/T}}{\sum_{k=1}^{N_s} e^{o_k(x_i)/T}} \tag{7}$$

实验设计

dataset

source domain：

- LFW
- CALFW
- CPLFW
- CFP-FP
- AgeDB-30
- IJB-C
- MS1M-RetinaFace

target domain：

- CASIA NIR-VIS 2.0
- QMUL-SurvFace
- KidsFace

实验结果

Model	Target Domain 1	Source Domain							Training Time(days)
	KidsFace -Test	LFW	LFW BLUFR	CALFW	CPLFW	CFP-FP	AgeDB-30	IJB-C	
JT(Upper Bound)	90.239	99.75	99.84	95.98	92.88	98.11	98.03	95.19	2.4
BaseS	53.687	99.73	**99.84**	**95.98**	**92.63**	**98.11**	**98.03**	**95.02**	5.5
BaseT	70.872	90.42	25.02	74.36	64.48	68.08	66.20	35.48	0.3
FT	84.661	96.08	60.62	84.43	75.53	79.47	79.40	78.84	0.16
FAwF	**86.846**	**99.75**	99.79	95.95	92.19	97.82	97.91	94.37	**0.1**

实验结果

Method	Target Domain 1	Source Domain								
	KidsFace-Test	LFW	LFW BLUFR	CALFW	CPLFW	CFP-FP	AgeDB-30	IJB-C	MF1 Rank 1	MF1 Veri.
Contrastive	83.753	99.22	95.53	93.15	87.75	94.70	94.06	83.02	83.815	85.484
Trplet	85.057	99.20	94.22	92.96	86.25	93.48	92.18	81.80	85.823	86.676
LwF [30]	84.977	95.83	52.80	84.63	73.96	77.15	79.71	59.04	40.145	37.355
iCaRL [32]	73.894	99.68	99.35	95.76	90.38	95.97	96.95	91.74	94.117	95.215
EEIL [33]	73.585	99.70	99.49	95.70	90.70	95.57	97.31	91.91	95.162	96.030
BiC [34]	81.722	99.45	98.72	95.33	87.65	91.54	95.95	79.69	90.681	92.508
FAwF	**86.846**	**99.75**	**99.79**	**95.95**	**92.19**	**97.82**	**97.91**	**94.37**	**96.899**	**97.123**

实验结果

Method	Target Domain 2	Target Domain 1	Source Domain						
	CASIA NIR-VIS 2.0	KidsFace-Test	LFW	LFW BLUFR	CALFW	CPLFW	CFP-FP	AgeDB-30	IJB-C
Contrastive	99.291	30.585	98.72	87.70	90.96	85.21	92.18	91.35	78.01
Trplet	99.267	57.864	98.93	92.48	92.11	84.23	90.72	91.01	71.57
LwF [30]	97.881	70.706	96.08	54.91	84.30	73.16	75.87	78.91	58.10
iCaRL [32]	95.586	72.425	99.68	99.52	95.61	90.63	95.84	97.18	91.20
EEIL [33]	93.862	70.412	99.72	99.45	95.68	90.20	95.41	97.21	89.83
BiC [34]	94.660	72.596	99.50	98.53	95.05	86.80	90.01	95.65	76.62
FAwF	**99.629**	**84.731**	**99.73**	**99.78**	**95.95**	**92.36**	**98.05**	**97.96**	**94.33**

实验结果

Method	Target Domain 3	Target Domain 2	Target Domain 1	Source Domain						
	QMUL-SurvFace	CASIA NIR -VIS 2.0	KidsFace -Test	LFW	LFW BLUFR	CALFW	CPLFW	CFP-FP	AgeDB-30	IJB-C
Contrastive	44.2	1.216	10.974	86.07	4.78	64.81	63.48	65.00	62.01	14.62
Trplet	50.0	1.006	17.139	91.32	20.24	70.93	69.10	72.60	64.43	24.22
LwF [30]	54.7	0.483	14.441	78.72	2.77	57.98	60.38	61.94	51.28	7.72
iCaRL [32]	56.5	54.977	63.519	99.60	99.29	95.30	87.21	88.48	96.36	86.12
EEIL [33]	57.5	54.131	64.492	99.63	99.27	95.53	87.05	88.50	96.56	86.30
BiC [34]	45.5	65.568	61.374	99.08	96.62	93.93	80.60	78.02	93.48	74.73
FAwF	**59.3**	**97.906**	**71.004**	**99.70**	**99.74**	**95.46**	**91.28**	**96.83**	**97.15**	**90.43**

实验结果

Ablation Study

PCE	MES	H&S	Target Domain 1	Source Domain								
			KidsFace-Test	LFW	LFW BLUFR	CALFW	CPLFW	CFP-FP	AgeDB-30	IJB-C	MF1 Rank 1	MF1 Veri.
-	-	H	86.039	98.85	86.32	94.21	85.30	87.87	93.70	85.88	80.466	82.978
✓	-	H	86.645	99.68	99.66	95.86	91.61	96.97	97.23	93.63	94.598	95.350
✓	✓	H	86.825	99.73	99.78	**95.96**	**92.53**	97.65	97.66	94.36	96.645	96.682
-	-	H&S	73.641	99.70	99.41	95.56	90.49	95.34	97.24	91.04	92.688	92.893
✓	-	H&S	**87.045**	99.73	99.69	95.76	91.76	97.45	97.28	93.93	95.294	96.385
✓	✓	H&S	86.846	**99.75**	**99.79**	95.95	92.19	**97.82**	**97.91**	**94.37**	**96.899**	**97.123**

思考与疑问

训练新模型时，蒸馏损失一项，为什么不只考虑source domain的sample带来的损失来维持source domain performance，将所有样本都考虑会不会影响到target domain performance？

$$L_d = -\frac{1}{M_t + N_s} \sum_{i=1}^{M_t+N_s} \sum_{j=1}^{N_s} \hat{\pi}_j(x_i) \log[\pi_j(x_i)] \tag{6}$$

$$\hat{\pi}_j(x_i) = \frac{e^{\hat{o}_j(x_i)/T}}{\sum_{k=1}^{N_s} e^{\hat{o}_k(x_i)/T}}, \quad \pi_j(x_i) = \frac{e^{o_j(x_i)/T}}{\sum_{k=1}^{N_s} e^{o_k(x_i)/T}} \tag{7}$$

Thanks!
谢谢