

Unified Representation Learning for Cross Model Compatibility

Chien-Yi Wang¹
chiwa@microsoft.com

Ya-Liang Chang²
yaliangchang@cmlab.csie.ntu.edu.tw

Shang-Ta Yang¹
shanya@microsoft.com

Dong Chen³
dch@microsoft.com

Shang-Hong Lai¹
shlai@microsoft.com

¹ Microsoft Cloud and AI
Taipei, Taiwan

² National Taiwan University
Taipei, Taiwan

³ Microsoft Research Asia
Beijing, China

Abstract

We propose a unified representation learning framework to address the Cross Model Compatibility (CMC) problem in the context of visual search applications. Cross compatibility between different embedding models enables the visual search systems to correctly recognize and retrieve identities without re-encoding user images, which are usually not available due to privacy concerns. While there are existing approaches to address CMC in face identification, they fail to work in a more challenging setting where the distributions of embedding models shift drastically. The proposed solution improves CMC performance by introducing a light-weight Residual Bottleneck Transformation (RBT) module and a new training scheme to optimize the embedding spaces. Extensive experiments demonstrate that our proposed solution outperforms previous approaches by a large margin for various challenging visual search scenarios of face recognition and person re-identification.

1 Introduction

Visual recognition and retrieval systems are widely deployed in our lives, such as frictionless physical access [19], missing person search [33], and global place recognition [18]. Most of these systems fall into the open-set visual recognition, which learns models to encode the images into unique embeddings in the high-dimensional vector space. Embeddings of the same class instances are clustered well in the embedding space, and accurate retrieval and recognition are achieved by finding the nearest neighbors in the space. Due to the fast progress of deep neural network architectures [9, 12, 13, 61], representation learning techniques [9, 21, 29] and large labeled training set [8, 52], diverse embedding models are deployed to meet the requirements for different scenarios. Moreover, embedding models with improved performance are released and updated continuously to achieve a better user experience.

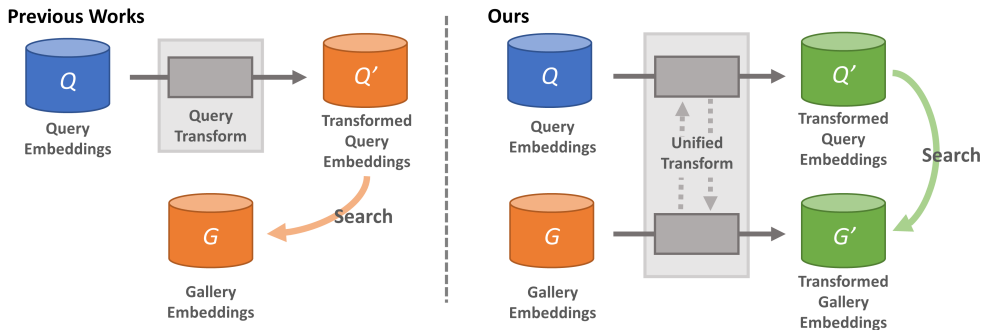


Figure 1: Cross Model Compatibility (CMC) issue takes place while query and gallery embeddings were encoded from different recognition models. Previous approaches address the issue by modeling the transformation between two embedding spaces directly, which has worse performance (Sec. 4.4) in many scenarios. Our proposed method aims to optimize a new unified embedding space, which achieves better cross-compatibility.

Common visual search applications like person re-identification [65] and face recognition [4] have registered *gallery* embeddings from a large number of identities, and test images are encoded from the same embedding model as *query* to perform recognition in the same embedding space. However, under some practical scenarios like upgrading the recognition model or searching across different device models, the system should be able to well address the Cross Model Compatibility (CMC) issue as the embedding spaces from different recognition models are not compatible with each other.

Re-encoding the gallery images with the same model seems to be a straightforward solution, but the original gallery images may not be stored in the system due to privacy concerns in the industry. Besides re-encoding the images, another feasible approach is to process *gallery* and *query* embeddings directly with another representation learning module to enable the compatibility. Chen *et al.* [4] took the first step in this direction to address CMC in face identification, and proposed R³AN to transform the *query* embeddings into the *gallery* embedding space while reconstructing realistic user face images at the middle stage. However, their approach only works when the two face embedding models are similar, but cannot generalize well in other practical scenarios where embedding models differ a lot.

In this work, we study extensively the relationship between embeddings across different embedding models and propose a unified representation learning framework to address the CMC problem, and it shows outstanding performance across various challenging scenarios. Inspired by ResNeXt [66], we propose the light-weight Residual Bottleneck Transformation (RBT) module to learn the embedding transformation very efficiently. Stacking fully connected layers would result in the heavy parameters and gradient vanishing problems during the training. RBT blocks mitigates these issues by skip-connection, channel down-scaling, and path splitting. Instead of transforming the *query* embeddings into the *gallery* embedding space, we propose to transform both *query* and *gallery* embeddings into a unified embedding space. We adopt similarity loss, dual classification loss, and KL loss in the framework to learn a new embedding space which clusters both the transformed *query* and *gallery* embeddings with low intra-subject variation as well as high inter-subject variation. Compared to previous approaches, our paradigm has one more degree of freedom which optimizes the unified embedding space to fit the CMC need better. Moreover, our unified framework gen-

eralizes well across various cross model scenarios and open-set visual recognition tasks.

The contributions of this work are summarized as follows:

- We formalize the Cross Model Compatibility (CMC) problem in the context of visual recognition and retrieval. This new problem aims to model the relationship between embedding spaces from different visual recognition models.
- We propose the light-weight Residual Bottleneck Transformation (RBT) module and a unified learning framework to overcome the CMC problem. The light-weight RBT module mitigates the model convergence issue in previous approaches, and the framework learns the dual transformation for *query* and *gallery* embeddings and achieves better compatibility in the new embedding space.
- The proposed RBT module and the learning framework demonstrate superior performance over previous approaches by up to 9.8% across challenging scenarios in face recognition and person re-identification tasks.

2 Related Works

2.1 Open-set visual recognition

Face recognition [1, 19], person re-identification [22, 35], and image retrieval [1, 25] are popular open-set visual recognition tasks. Deep neural networks (DNNs) are widely applied to learn embedding models that encode each image into an embedding vector. Open-set recognition and retrieval are achieved by computing distances between embedding vectors in the learned embedding space. Some methods train an embedding model by leveraging close-set classification as a surrogate task with various forms of loss functions [1, 27, 29], while others apply metric learning to enforce affinity between embeddings [11, 21]. Those methods provided a basis to train a robust visual embedding model to embed the identity images into the representation vectors.

2.2 Learning across domains

To address distribution change or domain shift issue in computer vision tasks, many domain adaptation [28] techniques are proposed to adapt the output distribution from multiple different modalities. Heterogeneous face recognition (HFR) [20] is one of them which aims to match face images acquired from different sources (i.e., different devices, resolutions, or wavelengths) for identification or verification. Because of the domain discrepancy between input image sets, data synthesis [17, 23] and domain-invariant feature learning [10, 14, 30] techniques are applied to address the problem. Our CMC problem differs in that the discrepancy comes from the embeddings itself instead of input images. While there are many works in domain adaptation addressing the distribution shift between images from different domains, the distribution shift between embeddings from different visual recognition models is not well studied.

2.3 Compatible representation learning

Shen *et al.* [22] propose a backward-compatible representation learning technique to learn visual embedding that is compatible with previous embedding models. However, given a

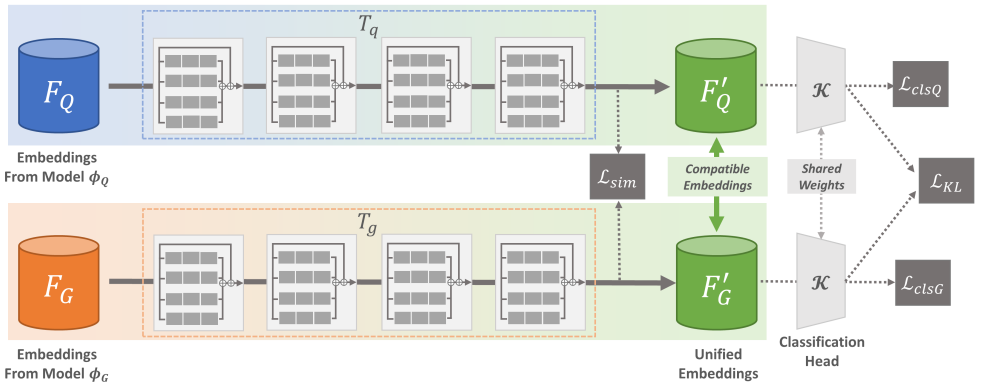


Figure 2: Overview of the proposed unified representation learning framework. Each of the non-linear transformation T_q and T_g between the embedding spaces is composed of four Residual Bottleneck Transformation (RBT) blocks. We optimize the network using three loss functions during the training process: Similarity loss \mathcal{L}_{sim} on the transformed embeddings F'_Q and F'_G , dual classification loss \mathcal{L}_{clsQ} , \mathcal{L}_{clsG} with the shared classifier, and KL-divergence loss \mathcal{L}_{KL} on the classifier outputs.

large variety of many different embedding models, it is impossible to train a new embedding model which is backward-compatible with every model. Moreover, the technique cannot apply to address the compatibility between deployed embedding models, which is the usual case in the industry. Chen *et al.* [2] raises the Cross Model Face Recognition (CMFR) problem, which is the sub-problem of CMC. The proposed R³AN approach is optimized for face identification and is hard to generalize to other visual search applications. We aim to formulate a more general CMC problem and propose a unified training framework to overcome the limitations in [2].

3 Proposed Method

3.1 Notations and Problem Formulation

We first define the notations to be used in the CMC problem. The CMC issue takes place while the *query* and *gallery* embeddings are encoded from different embedding models in the context of open-set recognition tasks. We denote the two embedding models as ϕ_Q and ϕ_G . For a group of N sample images, two sets of embeddings are encoded from these images: $F_Q = \{(x_i^q, y_i^q)\}_{i=1}^N$ and $F_G = \{(x_i^g, y_i^g)\}_{i=1}^N$, where $x_i^q \in \mathbb{R}^{d_q}$, $x_i^g \in \mathbb{R}^{d_g}$ denotes the embeddings of the i -th sample, belonging to the y_i -th identity class, and d_q and d_g are the corresponding dimension of embedding models. To address CMC, additional transformation T_q and T_g need to be learned and applied onto F_Q and F_G : $F'_Q = T_q(F_Q)$, $F'_G = T_g(F_G)$, and the resulting embedding sets F'_Q and F'_G are compatible with each other.

3.2 Residual Bottleneck Transformation (RBT) Module

To address CMC with high dimensional embedding sets, a common choice to model the relationship between embedding spaces is the non-linear mapping by multi-layer perceptron

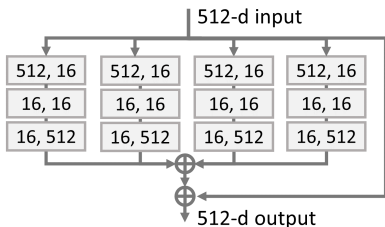


Figure 3: Network architecture of the proposed RBT module. Each small block has {FC, BatchNorm, ReLU}, and the number denotes the input and output dimensions.

Methods	T_q	T_g	Losses
MLP baseline	MLP	<i>Identity</i>	\mathcal{L}_{sim}
RBT baseline	RBT blocks	<i>Identity</i>	\mathcal{L}_{sim}
R ³ AN [10]	Decoder-Encoder	<i>Identity</i>	$\mathcal{L}_{sim}, \mathcal{L}_{img}^*$
Ours	RBT blocks	RBT blocks	$\mathcal{L}_{sim}, \mathcal{L}_{cls}, \mathcal{L}_{KL}$

Table 1: Comparison of different methods for addressing CMC, using the same annotation in the unified framework. *Identity* denotes the identity transformation. \mathcal{L}_{img}^* consists of the reconstruction and adversarial losses in [10] which require face input images for training.

(MLP). However, there are serious drawbacks with MLP: 1) As we show in Sec 4.4, building a transformation network with deeper MLP cannot lead to better performance. It is highly related to the gradient vanishing issue resulted from a large variation of the weight parameters [16]. 2) The number of the parameters and FLOPs (floating-points operations) will blow up while stacking high-dimensional transform through fully connected layers.

To overcome the above challenges, we propose the Residual Bottleneck Transformation (RBT) module. The detailed architecture of the module is shown in Fig 3. The RBT module exploits the split-transform-merge strategy, which is widely used in the backbone of Convolutional Neural Networks (CNNs) [24, 51], to save the parameters and operations while keeping the representation power. The input is split into four paths of low-dimensional embeddings (bottleneck), transformed by another module, and merged by concatenation. It also mitigates the gradient vanishing issue by the residual skip-connection [9]. We leverage the RBT module to build up our strong baseline and the unified learning framework.

3.3 Unified Representation Learning Framework

We propose a unified representation learning framework (Fig. 2) to learn better transformation T_q and T_g to address CMC. The motivation is to learn an unified embedding space which leverage information from both *query* and *gallery* embedding spaces. To optimize the compatibility between the transformed embedding sets F'_Q and F'_G , the unified embedding space needs to be discriminative enough to separate the identities. Firstly, we employ several RBT module blocks into T_q and T_g to encourage efficient non-linear embedding transformation. Secondly, the transformed embedding sets F'_Q and F'_G are passed into the shared classification head h to classify the identities. During the training stage, the closed-set identity classification is treated as a surrogate task that provides a supervision signal to improve generalization ability of the transformed embeddings. The classification head h is shared to ensure the compatibility between the two embedding spaces. Several loss functions are applied to train the network:

Similarity Loss. Given the sets of the transformed embedding $F'_Q = T_q(\{(x_i^q, y_i^q)\}_{i=1}^N)$ and $F'_G = T_g(\{(x_i^g, y_i^g)\}_{i=1}^N)$, the embeddings from the same input sample are enforced to be closed in the unified embedding space as they contain the similar semantic information. We apply

\mathcal{L}_2 losses as the similarity loss on the transformed embeddings:

$$\mathcal{L}_{sim} = \sum_{i=1}^N |T_q(x_i^q) - T_g(x_i^g)|_2 \quad (1)$$

Dual Classification Loss. The classification head \mathcal{K} classify the identities of the transformed embeddings to provide the supervision signal. By sharing the classification head, it enforces the embedding spaces of F_Q' and F_G' to be aligned with each other. The choice of the head \mathcal{K} (e.g. Softmax, Cosine [27], AM-Softmax [26], Arcface [8]) depends on the visual recognition task, and we denote the weights and biases in the fully connected head as \mathbf{W} and \mathbf{b} . The dual classification loss is calculated as follows:

$$\mathcal{L}_{cls} = \mathcal{L}_{clsQ} + \mathcal{L}_{clsG} = -\frac{1}{N} \sum_{i=1}^N (\log \mathcal{K}(T_q(x_i^q), y_i^q, \mathbf{W}, \mathbf{b}) + \log \mathcal{K}(T_g(x_i^g), y_i^g, \mathbf{W}, \mathbf{b})) \quad (2)$$

KL-divergence Loss. We also penalize the KL-divergence of the classifier output probabilities between transformed *query* and *gallery* embedding from the same input sample, and the KL-divergence loss is calculated as follows:

$$\mathcal{L}_{KL} = \sum_{i=1}^N KL((\mathcal{K}(T_q(x_i^q), y_i, \mathbf{W}, \mathbf{b}), \mathcal{K}(T_g(x_i^g), y_i, \mathbf{W}, \mathbf{b})) \quad (3)$$

The total loss for the proposed unified training framework is:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{sim} + \lambda_2 \mathcal{L}_{cls} + \lambda_3 \mathcal{L}_{KL} \quad (4)$$

where λ_1 , λ_2 , and λ_3 are weights for the \mathcal{L}_{sim} , \mathcal{L}_{cls} , and \mathcal{L}_{KL} , respectively. In all the experiments, we empirically set $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 0.25$.

4 Experiments

To verify the effectiveness of our proposed unified learning framework for cross model compatibility, we design a series of experiments with different cross model scenarios. We assess our proposed RBT module and the unified representation learning framework for the face identification problem and compare the compatibility performance with other approaches. We prepare and train various commonly used face embedding models to provide challenging cross model recognition tasks. Then we extract embeddings of the samples in the training dataset from prior face embedding models to address CMC between these embedding sets.

4.1 Prior Face Embedding Models

To have a holistic comparison of CMC and verify the generalization ability of our proposed framework, we prepare various prior embedding models with different network backbones and different training loss for identity classification. In real world applications, those in-play embedding models were not all using the same training loss, which would lead to very different embedding distributions, and the modeling of the relationship between embedding spaces could be more challenging. In total there are six prior face embedding models (Table 2), which were all trained on the same training dataset MS1M-retinaface [8], and were

Table 2: Details of prior face embedding models. Network: backbone architecture of the model. Name: abbreviation for the model. Loss: training loss for identity classification. FLOPs: floating-point operations, in # of multiply-adds. Top-1 Acc: top-1 identification accuracy on MegaFace [15] with 1M distractors.

Network	Name	Loss	FLOPs	Top-1
ResNet-100 [10]	R100	ArcFace	24G	98.90
MobileFaceNet [9]	Mb	ArcFace	933M	95.50
DenseNet-290 [12]	Dns	AM-Softmax	25G	98.63
ProxylessNAS [11]	Pxy	AM-Softmax	639M	96.04
ResNet-100 [10]	R100s	Softmax	24G	87.83
ResNet-50 [10]	R50s	Softmax	12.6G	89.84

Table 3: Comparison between different model architectures in $R100 \rightarrow Mb$ scenario. MLP(n) represents the MLP model with n hidden layers. RBT(n) represents the model with n RBT blocks.

Models	params	FLOPs	Top-1
MLP(1)	0.53M	0.52M	96.19
MLP(2)	0.79M	0.79M	96.30
MLP(3)	1.05M	1.05M	95.13
MLP(4)	1.31M	1.31M	71.27
RBT(1)	0.33M	0.33M	96.70
RBT(2)	0.40M	0.40M	96.74
RBT(3)	0.47M	0.47M	96.78
RBT(4)	0.54M	0.54M	96.93
R ³ AN [13]	6.64M	194.48M	97.04
Ours	1.08M	1.08M	97.58

evaluated on the MegaFace [15] (challenge 1 with FaceScrub as probe set). The MS1M-retinaface dataset is based on MS1M dataset [8] and refined by RetinaFace [9] face detector. The output dimension is 512 for all face embedding models.

4.2 Baseline approaches

The naive baseline approach is to learn the transformation T_q between *query* and *gallery* embedding models with multi-layer perceptron (MLP) using the \mathcal{L}_2 similarity loss. We denoted this approach as the **MLP baseline**. As discussed in Sec. 3.2, we proposed the Residual Bottleneck Transformation (RBT) module to overcome the limitation while learning the high-dimensional embedding mapping with MLP. Therefore, we also build a strong **RBT baseline**, which replace the MLP with RBT blocks, while using the same \mathcal{L}_2 similarity loss. All the approaches for comparison are summarized under the same framework in Table 1.

4.3 Implementation details

We implement the baselines and proposed training framework using Pytorch. The MLP baseline in all experiments is built by two hidden layers, which reaches the best performance, with 512-dimension output in each hidden layer. In the RBT baseline and the proposed solution, we employ four blocks of RBT in T_q and T_g . We choose Arcface [9] as the classification head \mathcal{K} in the face identification experiments with the margin parameters $s = 64$ and $m = 0.5$. The baseline models and the proposed framework are all trained with learning rate starting from 0.1 and divided by 10 after 20 and 25 epochs, and terminate the training after 30 epochs. We also re-implement and train the previous work R³AN [13] with the same protocol described in the paper.

4.4 Experimental results

Evaluation protocol. In the following experimental results, we use the face identification task MegaFace [15] challenge 1 with facescrub as the probe and 1M distractors to evaluate the compatibility between prior embedding models. The experiment is denoted as $M_q \rightarrow M_g$

Table 4: CMC results on embedding models with similar distribution. (M_q, M_g) denotes the original identification accuracy of the (*query*, *gallery*) embedding models.

Methods	R100→Mb (98.90, 95.50)	Dns→Pxy (98.63, 96.04)	R100s→R50s (89.84, 87.83)	Mb→R100 (95.50, 98.90)	Pxy→Dns (96.04, 98.63)	R50s→R100s (87.83, 89.84)
MLP baseline	96.30	91.47	87.42	92.98	93.75	87.30
RBT baseline	96.93	94.57	87.44	96.95	95.00	87.94
R ³ AN [□]	97.04	94.97	86.55	96.75	96.10	86.21
Ours	97.58	97.27	91.23	97.26	96.83	91.01

Table 5: CMC results on embedding models with large distribution shift. (M_q, M_g) denotes the original identification accuracy of the (*query*, *gallery*) embedding models.

Methods	R100→R50s (98.90, 87.83)	R50s→Dns (87.83, 98.63)	Dns→Mb (98.63, 95.50)	Mb→R100s (95.50, 89.84)	R100s→Pxy (89.84, 96.04)	Pxy→R100 (96.04, 98.90)
MLP baseline	85.84	58.26	94.57	86.21	75.41	84.12
RBT baseline	87.85	87.86	96.56	86.26	87.07	96.43
R ³ AN [□]	88.90	87.90	96.68	86.87	86.04	82.34
Ours	95.09	92.67	97.14	92.40	93.07	96.59

Methods	R50s→R100 (87.83, 98.90)	Dns→R50s (98.63, 87.83)	Mb→Dns (95.50, 98.63)	R100s→Mb (89.84, 95.50)	Pxy→R100s (96.04, 89.84)	R100→Pxy (98.90, 96.04)
MLP baseline	89.76	81.68	83.41	85.48	77.27	89.28
RBT baseline	90.15	86.84	93.54	87.86	84.46	96.45
R ³ AN [□]	87.44	85.00	85.36	87.81	82.72	86.13
Ours	94.58	93.56	96.65	92.75	92.80	97.19

if we use transformed embeddings from M_q as probe and transformed embeddings from M_g as distractors in the evaluation. The top-1 face identification accuracy is reported in the table.

Effects on different architectures. In Table 3, we demonstrate the difference in model parameters, FLOPs and CMC performance under the $R100 \rightarrow Mb$ scenario. We observe that we cannot build a deep MLP model as the CMC performance drop significantly with increasing hidden layers. It demonstrates that the MLP model with more parameters suffers from gradient vanishing issue. Note that the MLP baseline reported in R³AN [□] is much lower than expected as it built the MLP with the same parameters as the R³AN model. The proposed RBT module mitigates the issues of MLP, and performs better as we increase the number of blocks. In the following experiments, we use MLP with two hidden layers and RBT with four blocks as our strong baselines. Our proposed RBT baseline and unified framework both exhibit comparable results with R³AN [□], but with only **16.27%** and **0.56%** of the parameters and FLOPs, respectively.

Comparisons of different CMC approaches. Table 4 shows comparisons of CMC results, under the scenarios where two prior embedding models have similar distributions, as they were trained with the same classification loss. By optimizing the unified embedding space for two embedding sets, our proposed framework exhibit greater ability to address CMC than other approaches. Baseline and R³AN [□] approaches, which only transform one side of the embeddings, can achieve good compatibility results as the distribution changes between embedding models are fairly mild. Notably, in every case, the RBT baseline achieves comparable results with R³AN [□], which suggests that we do not need complex network architecture to model the distribution shift between embedding sets.

We further evaluate the CMC approaches with scenarios where two prior embedding models were trained with different classification losses. The comparison results are shown

in Table 5. We observe that the training of R³AN [2] is more unstable than the previous experiments, and requires more tweaking to achieve comparable results. MLP baseline cannot achieve comparable results under these scenarios, as it suffers from the network capacity and training instability. Under such challenging scenarios, our solution performs significantly better than RBT baseline and R³AN [2] by a large margin, which suggests that the general CMC is better addressed by optimizing another embedding space to adapt embeddings from different distributions. In the scenarios of (Dns, Mb) and (Pxy, R100), the RBT baseline can still produce comparable results, which indicates that embedding distributions supported by AM-softmax [26] and Arcface [4] do not differ a lot.

Effects on the training losses. We conduct ablation studies on the proposed approach with different training loss combinations and summarize the results in Table 6. The classification losses by the shared classifier contribute the most for optimizing the unified embedding space, as it encourages the transformed embeddings to be more discriminative on the identities. The similarity loss and the KL loss also improve the identification accuracy by 0.32% and 0.25%. Lastly, the proposed method reaches its summit by employing all the losses.

Transformation cost. One may raise concerns about the proposed approach as it requires additional transformation on the *gallery* embedding sets. In practical applications, we only need *one-time* transformation on the existed gallery embedding sets. By leveraging the lightweight RBT modules with only 0.54M FLOPs, the transformation on 1M gallery embeddings only takes 0.06 second on a single TITAN RTX GPU, which is very efficient. Moreover, for each on-the-fly *query* transformation, it saves more than 99% of floating points operations compared with R³AN [2].

4.5 Experiments on Person Re-identification

To verify the effectiveness of the unified framework, we validate the proposed method on the person re-identification task using the Market-1501 [62] benchmark. There are 751 and 750 identities in the training and testing dataset, respectively. We choose three pretrained person embedding models [64] in our CMC experiments: OSNet-1.0 (OS100), OSNet-0.25 (OS25), and MobilenetV2 (Mb). These models differ in the network backbone and the embedding dimensions, that the OSNet has 512-dim, and MobilenetV2 has 1280-dim. In our experiments, to accommodate different input dimensions, the down-scaling dimension in each path of the RBT module is set to $\frac{d_{in}}{32}$. During the training stage of the unified framework, the classification head \mathcal{K} employs Softmax with label smoothing [65], which is commonly used in the training of person re-identification models. Search mean average precision (mean AP) is used as the evaluation metric.

Table 7 demonstrates that our proposed RBT module and the unified learning framework can address the CMC issue in the person re-identification task. Besides, our method performs significantly better than MLP and RBT baselines for all scenario, which suggests that the unified learning scheme can better optimize the compatibility between different embedding models. In the future works, larger person re-identification datasets can be leveraged for further improvements on CMC.

Table 6: Ablation experiments on the training losses in the R100→Mb scenario.

\mathcal{L}_{cls}	\mathcal{L}_{sim}	\mathcal{L}_{KL}	Top-1
✓			97.10
✓	✓		97.42
✓		✓	97.35
✓	✓	✓	97.58

Table 7: CMC results on person re-identification. Search mean average precision (mAP) is reported.

Methods	OS100 → OS25 (82.6, 75.0)	OS100 → Mb (82.6, 67.3)	OS25 → Mb (75.0, 67.3)
MLP	67.1	46.7	38.4
RBT	68.0	58.9	54.1
Ours	74.1	67.3	62.9

Methods	OS25 → OS100 (75.0, 82.6)	Mb → OS100 (67.3, 82.6)	Mb → OS25 (67.3, 75.0)
MLP	66.1	57.1	51.3
RBT	67.9	59.4	54.9
Ours	76.3	65.3	59.1

5 Conclusions

We have presented a unified learning framework for addressing cross model compatibility (CMC) problem in the context of visual search and recognition applications. Our framework robustly optimizes a unified embedding space that adapts embedding distributions from two different embedding models to address CMC. Besides, we proposed a light-weight RBT embedding transformation module to facilitate the training stability and inference efficiency. Based on experimental results, we show that the proposed module and unified framework performs significantly better than previous approaches by a large margin under challenging scenarios in face identification and person re-identification.

References

- [1] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.
- [2] Ken Chen, Yichao Wu, Haoyu Qin, Ding Liang, Xuebo Liu, and Junjie Yan. R3 adversarial network for cross model face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9868–9876, 2019.
- [3] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenet: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018.
- [4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.
- [5] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [6] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

- [7] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016.
- [8] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [10] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1761–1773, 2018.
- [11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *ArXiv*, abs/1703.07737, 2017.
- [12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [14] Meina Kan, Shiguang Shan, and Xilin Chen. Multi-view deep network for cross-view classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4847–4855, 2016.
- [15] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.
- [16] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in neural information processing systems*, pages 971–980, 2017.
- [17] José Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6628–6637, 2017.
- [18] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2015.
- [19] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE.

- [20] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, Xueming Li, Chen Change Loy, and Xiaogang Wang. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution. *Image and Vision Computing*, 56:28–48, 2016.
- [21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [22] Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. *arXiv preprint arXiv:2003.11942*, 2020.
- [23] Lingxiao Song, Man Zhang, Xiang Wu, and Ran He. Adversarial discriminative heterogeneous face recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [25] Giorgos Tolias, Ronan Sifre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.
- [26] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.
- [27] Hai Jun Wang, Yitong Wang, Zuo-Feng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wenyu Liu. Cosface: Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.
- [28] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.
- [29] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.
- [30] Xiang Wu, Lingxiao Song, Ran He, and Tieniu Tan. Coupled deep learning for heterogeneous face recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [31] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [32] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.
- [33] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

-
- [34] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv preprint arXiv:1910.10093*, 2019.
 - [35] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3702–3712, 2019.