Residual Compensation Networks for Heterogeneous Face Recognition

Zhongying Deng,[*] Xiaojiang Peng,[*] Yu Qiao[†]
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
Shenzhen, Guangdong Province, China 518055
{zy.deng1, xj.peng, yu.qiao}@siat.ac.cn

Abstract

Heterogeneous Face Recognition (HFR) is a challenging task due to large modality discrepancy as well as insufficient training images in certain modalities. In this paper, we propose a new two-branch network architecture, termed as Residual Compensation Networks (RCN), to learn separated features for different modalities in HFR. The RCN incorporates a residual compensation (RC) module and a modality discrepancy loss (MD loss) into traditional convolutional neural networks. The RC module reduces modal discrepancy by adding compensation to one of the modalities so that its representation can be close to the other modality. The MD loss alleviates modal discrepancy by minimizing the cosine distance between different modalities. In addition, we explore different architectures and positions for the RC module, and evaluate different transfer learning strategies for HFR. Extensive experiments on IIIT-D Viewed Sketch, Forensic Sketch, CASIA NIR-VIS 2.0 and CUHK NIR-VIS show that our RCN outperforms other state-of-the-art methods significantly.

Introduction

Heterogeneous face recognition (HFR) mainly focuses on identifying a person from face images with different modalities such as photos versus sketches or near-infrared (NIR) versus visual (VIS) images. Owing to its promising applications in surveillance and law enforcement agencies, HFR has attracted increasing attention recently (Wu et al. 2017; Song et al. 2017; He et al. 2017). Though great progress has been made, it is still a challenging problem due to insufficient training samples and the large modal discrepancy.

With limited training data, many early works (Liao et al. 2009; Klare, Li, and Jain 2011; Li et al. 2016) address heterogeneous face recognition based on hand-crafted features. To reduce modal discrepancy, they mainly project face features of different modalities into latent common subspace. Nonetheless, hand-crafted features based methods gradually reach a bottleneck because hand-crafted features are with limited representation capability.

Recently, new breakthrough marked by deep Convolutional Neural Networks (CNN) has been made in various visual tasks including face recognition (Sun, Wang, and Tang 2014b; 2014a; Wen et al. 2016). Compared to hand-crafted features, CNN learned from large scale datasets extracts more discriminative features which describe the highly nonlinear relationship of different modalities. Therefore, some researchers introduce CNN to deal with the HFR problem and achieve impressive performance (Hu et al. 2017; Wu et al. 2017; Sarfraz and Stiefelhagen 2015; He et al. 2017). Though CNN has shown superior performance compared to traditional features, it may easily over-fit on small scale HFR datasets with naive training schemes. In addition, extracting features with a generic face CNN model for both NIR/Sketch and VIS images may still suffer modal discrepancy since all the parameters are shared while the inputs are of different modalities.

To prevent the target CNN model from over-fitting, Hu *et al*. (Hu et al. 2018) propose a synthetic data augmentation method to exponentially enlarge HFR datasets. Wu *et al*. (Wu et al. 2017) propose a coupled deep learning (CDL) approach which leverages a relevance constraint and a triplet loss to alleviate over-fitting and reduce modal difference. Several other methods utilize metric learning and generative networks for the two issues of HFR (Saxena and Verbeek 2016; He et al. 2017). Overall, these methods use a generic face CNN model to learn modality-invariant features which can be seen as to project different modal images into a common subspace.

In this paper, instead of using a generic CNN model, we propose a new two-branch network architecture, termed as Residual Compensation Networks (RCN), to learn separated features for different modalities in HFR. In our RCN, based on a well-trained backbone face CNN model from one modality with rich training data, we add a Residual Compensation (RC) module for the other modality, and tune it with an extra Modal Discrepancy loss (MD loss). The RCN tackles over-fitting by fixing the backbone face CNN model and tuning a light RC module for the other modality, and alleviates modal discrepancy by the RC module and the MD loss. Take the NIR-VIS task for example, we first train a VIS face CNN model on public available large VIS datasets, and then fix it for the VIS branch while add a RC module after the feature layer for the NIR branch, and finally tune the RC module with the MD loss and the cross-entropy loss on paired NIR-VIS face images with the same identities.

With a two-branch architecture, our RCN owns several

[*]Z. Deng and X. Peng contributed equally.
[†]Corresponding author.

advantages as follows: (i) It keeps the exact same representation capability for the VIS modality which is critical for some HFR cases like the Forensic Sketch dataset where VIS yields the enlarged gallery set. (ii) By tuning a light RC module with few parameters, it reduces the over-fitting efficiently. (iii) It also benefits from large scale paired image inputs which alleviates over-fitting. (iv) It reduces modal discrepancy with a easy-to-implement cosine based MD loss while keeps inter-identity discriminative power by standard cross-entropy loss.

In summary, we propose a novel two-branch modal-invariant deep learning framework for HFR with state-of-the-art results on several datasets. Our contributions are as follows:

- We propose the Residual Compensation Convolutional Neural Network to alleviate over-fitting and reduce modality discrepancy simultaneously for HFR, which can extend to other cross domain tasks instead of HFR.
- We design an easy-to-implement Modality Discrepancy loss (MD loss) to efficiently reduce modal discrepancy.
- Our RCN achieves the state-of-the-art performance on four popular HFR datasets, namely 90.34% on IIIT-D Viewed Sketch, 62.26% on Forensic Sketch, 99.32% on CASIA NIR-VIS 2.0 and 99.44% on CUHK NIR-VIS.

Residual Compensation Networks

To reduce the modal discrepancy, we propose a novel residual compensation (RC) module and a modality discrepancy loss (MD loss). In this section, we first present the overview of our RCN architecture. Then we elaborate the design of RC module and MD loss, followed by some discussions.

Overview of RCN

Figure 1 shows the architecture of our RCN. RCN takes as input an image pair of the same subject but with different modalities. Then the image pair is processed by two branches of RCN, we assume the right branch for VIS images and the left branch for NIR/Sketch images. The backbone CNN of both branches is a variation of ResNet (He et al. 2016), i.e. ResNet-10 in the right of Figure 1. The ResNet-10 consists of a FC layer and 10 convolution layers including 2 ResNet blocks whose block number is 1 and 2 respectively. The 128-d output vector from FC is considered as the face representation of VIS branch. For the NIR/Sketch branch, this vector is fed into the proposed RC module which constitutes a FC layer followed by a PReLU layer (He et al. 2015). The output of RC module is the final representation of a NIR/Sketch face. The whole network can be trained end-to-end with the joint supervision of cross entropy loss and the proposed MD loss. For testing, we use the left and right branch to extract the features of NIR/Sketch and VIS face images respectively.

Residual Compensation Module

To utilize a well-trained CNN model on VIS data efficiently, we propose a novel residual compensation module and add it to the NIR/Sketch branch.

Assume that the pre-trained CNN on large VIS face datasets is $f_\theta(*)$ with parameters θ, a VIS face image I_i^v and a NIR/Sketch face image I_i^n with the same identity, we can extract face features as $\mathbf{x}_i^v = f_\theta(\mathrm{I}_i^v)$ and $\mathbf{x}_i^n = f_\theta(\mathrm{I}_i^n)$ for both images. Note that $f_\theta(*)$ is learned on VIS face datasets, it is suitable to extract the discriminative feature for I_i^v. However, using $f_\theta(*)$ to extract the feature of I_i^n may result in a bad face representation since the distribution of NIR/Sketch images and VIS images are quite different. In other words, the pre-trained CNN may bring modal discrepancy between $\mathbf{x}_i^v$ and $\mathbf{x}_i^n$.

Since outputs of the pre-trained CNN $\mathbf{x}_i^v$ and $\mathbf{x}_i^n$ are of the same identity, they should be associated with corresponding inherent hidden component $\mathbf{x}_i$. Suppose that there are transmission functions φ_n and φ_v so that

$$\mathbf{x}_i^v = \varphi_v(\mathbf{x}_i), \mathbf{x}_i^n = \varphi_n(\mathbf{x}_i). \tag{1}$$

We denote $\widetilde{\varphi}_n$ as an approximation inverse function such that $\mathbf{x}_i \approx \widetilde{\varphi}_n(\mathbf{x}_i^n)$. Then the difference between $\mathbf{x}_i^v$ and $\mathbf{x}_i^n$ is

$$\begin{aligned} \mathbf{x}_i^v - \mathbf{x}_i^n &\approx \varphi_v(\widetilde{\varphi}_n(\mathbf{x}_i^n)) - \mathbf{x}_i^n \\ &\approx \phi(\mathbf{x}_i^n), \end{aligned} \tag{2}$$

where $\phi(\mathbf{x}_i^n) = \varphi_v(\widetilde{\varphi}_n(\mathbf{x}_i^n)) - \mathbf{x}_i^n$. Eq.(2) shows that the modal gap between $\mathbf{x}_i^v$ and $\mathbf{x}_i^n$ can be approximatively modeled by a residual module, i.e.

$$\mathbf{x}_i^v \approx \mathbf{x}_i^n + \phi(\mathbf{x}_i^n). \tag{3}$$

In practice, $\mathbf{x}_i^v$ and $\mathbf{x}_i^n$ come from the FC features of two modalities of the same person. Since the FC features mainly encode facial identity information, it is expected that $\mathbf{x}_i^v$ and $\mathbf{x}_i^n$ are close to each other.

To reduce the modal discrepancy, we argue that the gap between the *desired* features $\hat{\mathbf{x}}_i^v = \mathbf{x}_i^v$ and $\hat{\mathbf{x}}_i^n$ can be reduced by compensating $\mathbf{x}_i^n$ with g_τ, where g_τ is a mapping function with parameters τ. Specifically, we add a residual compensation (RC) module g_τ into the NIR/Sketch branch of the pre-trained CNN $f_\theta(*)$ to make $\hat{\mathbf{x}}_i^n$ approach $\hat{\mathbf{x}}_i^v$, i.e. $\hat{\mathbf{x}}_i^n = \mathbf{x}_i^n + g_\tau(\mathbf{x}_i^n) \approx \hat{\mathbf{x}}_i^v$. The objective of the RC module can be formulated as

$$\begin{aligned} \min_{\tau} \quad & \sum_i diff(\hat{\mathbf{x}}_i^v, \hat{\mathbf{x}}_i^n), \\ s.t. \quad & \hat{\mathbf{x}}_i^v = \mathbf{x}_i^v = f_\theta(\mathrm{I}_i^v), \\ & \hat{\mathbf{x}}_i^n = \mathbf{x}_i^n + g_\tau(\mathbf{x}_i^n) = f_\theta(\mathrm{I}_i^n) + g_\tau(f_\theta(\mathrm{I}_i^n)), \end{aligned} \tag{4}$$

where $diff(*,*)$ is a function that measures the difference of two inputs. If we fine-tune the backbone model as well, we can rewrite Eq.(4) by replacing f_θ as $f_{\theta+\Delta}$ where Δ denotes the weight changes of the pre-trained model. To minimize the difference of $\hat{\mathbf{x}}_i^v$ and $\hat{\mathbf{x}}_i^n$, we further propose the modality discrepancy loss.

Modality Discrepancy Loss

Considering the fact that we usually use cosine similarity to measure the difference of two face images, we can use the cosine distance as $diff(*,*)$ in Eq. (4). To this end, we

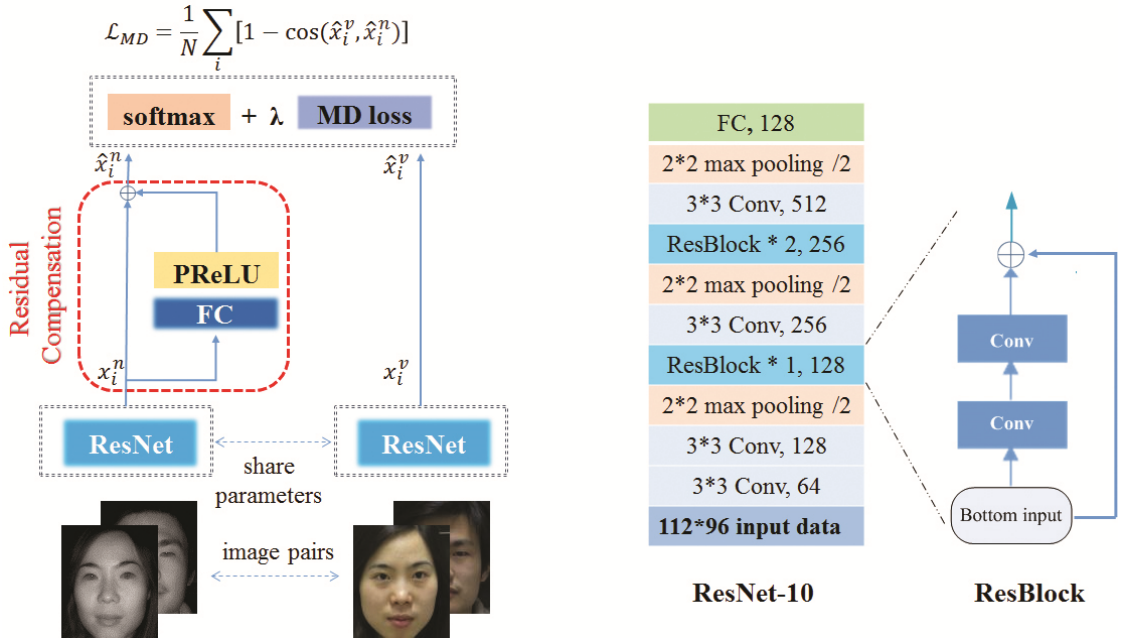

Figure 1: The pipeline of the our RCN. Left: the two-branch architecture of RCN. Right: the backbone model of RCN.

define the modality discrepancy loss (MD loss) between ($\hat{\mathbf{x}}_i^v$, $\hat{\mathbf{x}}_i^n$) as follows,

$$\mathcal{L}_{MD} = \frac{1}{N}\sum_{i=1}^{N}(1 - \cos(\hat{\mathbf{x}}_i^v, \hat{\mathbf{x}}_i^n)) \tag{5}$$

where $\cos(*,*)$ is the cosine similarity of two inputs and N is the total number of image pairs. It is obvious that the optimization of $\mathcal{L}_{MD}$ would force two face representations to be similar. The total loss can be formulated as

$$\mathcal{L} = \mathcal{L}_S + \lambda\mathcal{L}_{MD}, \tag{6}$$

where $\mathcal{L}_S$ is the cross entropy loss of face classification and λ is a hyper parameter to trade off these two terms. We define a deep neural network with residual compensation module and modality discrepancy loss as the Residual Compensation Networks (RCN).

Discussions

The architecture of RC module. By default, we implement g_τ with a fully-connected (FC) layer followed by a non-linear activation function and we refer it as standard RC module. This implementation owes to the following consideration: First, we only use a single FC layer because there are large amount of parameters in FC layer. More FC layers may lead to over-fitting easily. In practice, we add a dropout layer to further reduce over-fitting in training phase. Second, the non-linear activation function is used to improve the representation capability because i) the residual compensation is not necessary a linear mapping and ii) the relation between VIS and NIR/Sketch signals is highly non-linear. In theory, we can add the RC module in any layer instead of the FC layer as in (Rebuffi, Bilen, and Vedaldi 2018). We conduct a comprehensive evaluation to examine different setups of RC module about this issue in experiments.

Comparison to other losses for deep face recognition. Compared to contrastive loss (Sun, Wang, and Tang 2014a), triplet loss (Schroff, Kalenichenko, and Philbin 2015) and center loss (Wen et al. 2016), the MD loss of RCN is efficient and easy to implement, which don't need to carefully constitute the negative pairs/triplets from the training set or to adjust hyper parameter of margin (compare to contrastive loss and triplet loss) and don't introduce additional parameters in loss design (center loss learns class centers). In addition, we use cosine distance as supervision since cosine similarity is adopted to measure two face images in test phase. We find Euclidean distance can be large at the beginning and dominate the whole training loss, which may be harmful for training and can lead to divergence. On the contrary, MD loss use cosine distance which never be larger than 1. As a result, MD loss is relatively small comparing to cross entropy loss at beginning and can only play an important role at later stage of training phase, which makes the training process more stable and easier to converge.

Relation to other residual block based works. He *et al*. (He et al. 2016) propose residual networks to ease the training of very deep networks. Rebuffi *et al*. (Rebuffi, Bilen, and Vedaldi 2018) also apply residual block to transfer learning by adding a small number of residual parameters to universal parametric family. Cao *et al*. (Cao et al. 2018) propose a Deep Residual EquivAriant Mapping (DREAM) block to map profile faces to frontal faces and get the pose-robust CNN. Our work is inspired by these residual block based works but with several differences. First, as far as we know, we are the first to exploit RC module for HFR task. Second, RCN has two branches to process features from two

modalities. To adapt the fixed pre-trained backbone to new NIR task and reduce model discrepancy, we add a new RC module to the NIR branch. Third, we introduce MD loss to explicitly constrain the modal similarity.

Experiments

In this section, we evaluate our proposed RCN on IIIT-D Viewed Sketch (Bhatt et al. 2012b), Forensic Sketch (Klare, Li, and Jain 2011), CASIA NIR-VIS 2.0 (Li et al. 2013) and CUHK NIR-VIS.

Dataset and Protocol

IIIT-D Viewed Sketch and Forensic Sketch are two widely-used sketch datasets in HFR. IIIT-D Viewed Sketch consists of 238 sketch-photo image pairs which are drawn by a professional sketch artist given corresponding digital images. Forensic Sketch contains 159 forensic sketches which are drawn by forensic sketch artists according to the verbal descriptions of the witnesses. The corresponding mug shot photos are later identified by the law enforcement agency. For IIIT-D Viewed Sketch, we take the same training and testing protocols as (Wu et al. 2017) where training set is with the 1,194 image pairs from CUFSF (Zhang, Wang, and Tang 2011) and the rank-1 identification accuracy on IIIT-D Viewed Sketch is reported. For Forensic Sketch, We follow the same partition protocol as (Peng et al. 2017; Klare, Li, and Jain 2011) which uses 106 subjects to train and 53 subjects to test. As in (Peng et al. 2017; Klare, Li, and Jain 2011), the gallery is extended by 10,000 face images of 10,000 persons from MegaFace Challenge (Kemelmacher-Shlizerman et al. 2016) to simulate real scenarios, and rank-50 accuracy of face identification is reported.

CASIA NIR-VIS 2.0 and CUHK NIR-VIS are two popular NIR-VIS datasets. CASIA NIR-VIS 2.0 contains 17,580 images of 725 subjects with variations in pose, age, resolution and expressions. As the standard evaluation protocols in (Li et al. 2013), we tune parameters on View 1 and report the rank-1 face identification accuracy and verification rate (VR)@false acceptance rate (FAR) on View 2. For CUHK VIS-NIR face dataset, there are 2,876 different subjects and each subject has only an infrared facial image and a visible counterpart. Following (Li et al. 2016), we use 1,438 infrared and visible image pairs as the training set and the remaining 1,438 pairs as the testing set.

Implementation Details

The face images are detected by MTCNN (Zhang et al. 2016) and five landmarks (nose, two eyes, mouse corners) of each face are obtained for alignment with similarity transform. In this way, we align and crop face images to 112×96. Some cropped examples are shown in Figure 2. After that, each pixel ([0,255], RGB channels) of the cropped face images is subtracted by 127.5 and then divided by 128.

We pre-train the backbone ResNet-10 on several web-collected data, including CASIA-WebFace (Yi et al. 2014), CACD2000 (Chen, Chen, and Hsu 2015), Celebrity+ (Liu et al. 2015), MSRA-CFW (Zhang et al. 2012), cleaned version of MS-Celeb-1M (Guo et al. 2016) provided by (Wu et al. 2015). We adopt joint supervision of cross-entropy loss and the center loss (Wen et al. 2016) to train the model. The pre-trained model gets 99.48% on LFW (Huang et al. 2007).

Then, we initialize the shared ResNet-10 with the pre-trained model and train our RCN with cross-entropy loss and the MD loss. We set the batch size to 128, i.e. 64 image pairs and initial learning rate to 0.01. To alleviate over-fitting, we freeze all convolutional layers of the pre-trained CNN and only train the FC layers and RC module. All experiments are carried out based on the Caffe (Jia et al. 2014).

Exploration of RC Module

In this section, we first compare our RCN to several baseline models, and then explore the architecture and position of RC module.

Comparison to baseline models. We conduct the comparison on CASIA NIR-VIS 2.0 and IIIT-D Viewed Sketch. We consider four comparable baseline models: (a) traditional single-branch model by fine-tuning the FC layer only, (b) traditional single-branch model with an extra FC layer and a PReLU layer, (c) two-branch model with an extra FC layer and a PReLU layer for NIR/Sketch branch, (d) two-branch model with an extra RC module for NIR/Sketch branch and fine-tuned with softmax loss. Model (a) and model (b) are two straightforward transfer learning methods. Model (c) is a comparable baseline to the RC module. Figure 3 illustrates these baseline architectures.

Table 1 shows the performance comparison between the baseline models and our RCN. We have several observations as follows.

- As expected, the pre-trained model gets the worst performance on both datasets which indicates there exists large modal discrepancy between VIS and NIR/Sketch face images.
- As a typical transfer learning scheme, fine-tuning all the layers (2nd row) of the pre-trained model improves the performance largely. Instead of fine-tuning all the layers, the baseline model (a) only fine-tunes the FC layer and obtains superior results especially on IIIT-D. We argue that fine-tuning all the layers may have higher risk from over-fitting than (a) since the HFR datasets are small.
- Adding a new FC layer is an alternative simple transfer learning scheme. As shown in the 4th row, the performance of this scheme w/o the PReLU layer on both datasets is inferior to baseline model (a). The non-linear PReLU layer improves performance slightly on CASIA NIR-VIS 2.0 while degrades significantly on IIIT-D. This can be explained by more parameters and the powerful non-linear operation lead to over-fitting easily on small datasets.
- As an architecture-comparable baseline model, model (c) is even worse than the model obtained by fine-tuning all layers. The PReLU in (c) makes the features from VIS and NIR/Sketch branch hard to match each other since the VIS branch does not have a PReLU layer.
- Adding a default RC module to the NIR/Sketch branch outperforms all other baseline models. Compared to

Figure 2: Cropped face images from (a) IIIT-D Viewed Sketch, (b) Forensic Sketch, (c) CASIA NIR-VIS 2.0 and (d) CUHK NIR-VIS.

Table 1: Comparison of different baseline models on IIIT-D Viewed Sketch and CASIA NIR-VIS 2.0. The numbers in "()" of the 4th and 5th rows denote the results with a PReLU layer after FC.

Model	IIIT-D Sketch Rank-1 (%)	CASIA NIR-VIS 2.0 Rank-1 (%)	CASIA NIR-VIS 2.0 VR@FAR=0.1% (%)
Pre-trained ResNet-10	52.10	86.58 ± 1.36	75.43 ± 1.88
Fine-tune all layers in ResNet-10	82.35	97.06 ± 0.43	96.43 ± 0.47
(a) Fine-tune FC only	86.97	98.76 ± 0.20	98.34 ± 0.26
(b) Fine-tune FC with an extra FC (PReLU)	84.8(79.41)	97.94 ± 0.29 (98.19 ± 0.36)	97.64 ± 0.42 (97.76 ± 0.25)
(c) An extra FC to NIR/Sketch branch (PReLU)	76.47(73.11)	97.33 ± 0.36 (95.77 ± 0.82)	97.15 ± 0.22 (95.35 ± 0.61)
(d) An extra RC module to NIR/Sketch branch	88.24	98.91 ± 0.17	98.52 ± 0.23
(d) + center loss	86.55	98.87 ± 0.25	98.36 ± 0.34
(d) + contrastive loss	88.66	98.72 ± 0.24	98.00 ± 0.37
RCN (RC + MD loss)	**90.34**	$\mathbf{99.32 \pm 0.15}$	$\mathbf{98.74 \pm 0.24}$

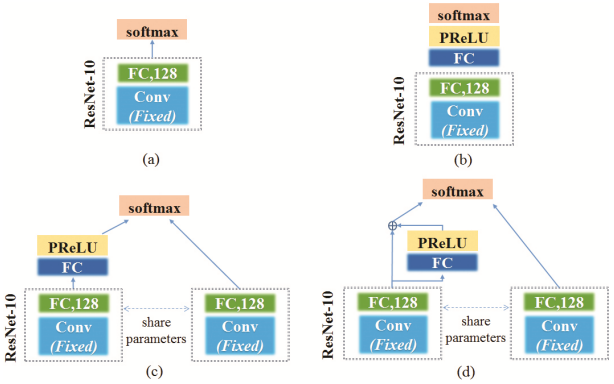

Figure 3: The architectures of four baseline models. (a) and (b) are single-branch while (c) and (d) are two-branch. (d) includes our default RC module.

model (c), our RC module benifits from two aspects: i) it keeps the main features of backbone networks and ii) it learns a powerful non-linear projection for the residual which compensates modality difference.

We compare our MD loss with two widely-used loss functions, namely center loss (7th row) and contrastive loss (8th row). Both loss functions are added to train model (d) with their best hyper-parameters from cross-validation. As shown in the last 3 rows, both center loss and contrastive loss slightly degrade or keep similar performance while our MD loss boosts performance on both datasets significantly. In our observation, the center loss fails to learn effective centers for subjects and contrastive loss suffers from finding effective negative/positive pairs with few samples. The values of both loss functions are very large compared to cross-entropy loss which makes training unstable. Overall, our RCN (RC+MD loss) outperforms the naive fine-tuning scheme (2nd row) in Rank-1 accuracy by 7.9% and 2.26% on IIIT-D and CASIA NIR-VIS 2.0, respectively.

Evaluation of RC implementation. We explore different RC designs with MD loss on IIIT-D and CASIA NIR-VIS 2.0 and compare their performance in Table 2.

First, we modify the standard RC module by removing the PReLU (1st row) layer. It decreases the Rank-1 accuracy of RCN by 2.1% on IIIT-D and 0.5% on CASIA NIR-VIS 2.0. This degradation verifies our consideration that non-linear mapping of g_τ can improve the representation capability. Second, we stack two standard RC modules (2nd row) to seek further improvements. It is slightly inferior to RCN, which can be explained by more parameters lead to over-fitting easily on small HFR datasets.

Similar to (Rebuffi, Bilen, and Vedaldi 2018), we design an alternative *conv* RC module (3rd row) which consists of a 3×3 convolutional layer and a PReLU layer, and add it to the first convolutional layer of the backbone network. As

Table 2: Performance of different RC implementations on IIIT-D Viewed Sketch and CASIA NIR-VIS 2.0.

RC module	IIIT-D Viewed Sketch Rank-1 (%)	CASIA NIR-VIS 2.0 Rank-1 (%)	CASIA NIR-VIS 2.0 VR@FAR=0.1% (%)
without PReLU	88.24	98.82 ± 0.35	98.37 ± 0.32
default RC $\times 2$	89.50	99.18 ± 0.18	98.70 ± 0.25
3*3 conv RC	88.66	98.93 ± 0.24	98.47 ± 0.21
default RC + 3*3 conv RC	87.82	99.31 ± 0.22	98.93 ± 0.29
RCN (RC + MD loss)	**90.34**	$\mathbf{99.32 \pm 0.15}$	$\mathbf{98.74 \pm 0.24}$

shown in Table 2, the *conv* RC is inferior to the default RC of our RCN especially on IIIT-D. Keeping both the default RC module and the *conv* RC module (4th row) gets comparable performance as our RCN on CASIA NIR-VIS 2.0 but degrades 2.52% on IIIT-D. We believe the difference is caused by the limited number of training samples of those two datasets.

Evaluation of RC position. Considering that we only test the *conv* RC module in the first convolutional layer in Table 2, we further evaluate the position of the *conv* RC module on CASIA NIR-VIS 2.0 with MD loss. Figure 4 shows the comparison in Rank-1 accuracy. As shown in Figure 4, adding RC module to the last convolutional layer or FC layer (default) achieves the best performance, which suggests that compensating high-level features is more effective. Adding RC to all layers degrades the performance of RCN from 99.32% to 98.81% which can be explained by too many parameters lead to over-fit easily.

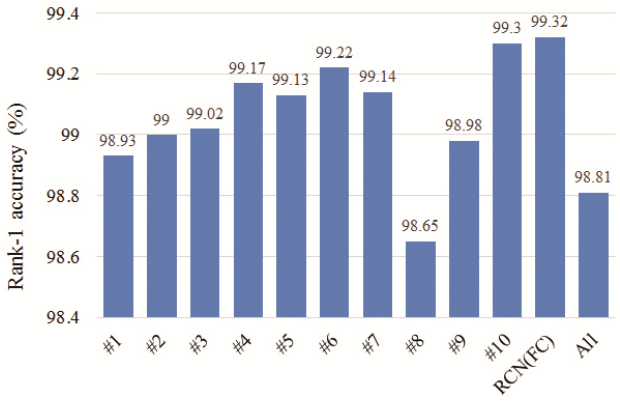

Figure 4: Evaluation of RC position on CASIA NIR-VIS 2.0. $\#k(k = 1, 2, ..., 10)$ means we add RC module to the k-th convolutional layer. 'All' means adding RC module to all layers.

Exploration of MD loss

We evaluate the hyper parameter λ in Eq. (6) on the IIIT-D Viewed Sketch and CASIA NIR-VIS 2.0. The results are shown in Figure 5.

From Figure 5, we can observe that the MD loss obtains superior performance to only using cross entropy loss (i.e. λ=0) with a wide range of λ on both datasets. The rank-1 accuracy is significantly increased with $\lambda = 0.4$ on IIIT-D and $\lambda = 0.8$ on CASIA NIR-VIS 2.0, which manifests the effectiveness of MD loss. In addition, the MD loss has the comparable performance to using cross entropy loss alone even with large λ (i.e. λ=1) on CASIA NIR-VIS 2.0. While a very high ratio of MD loss might be harmful for discriminability between classes and leads to degradation of performance.

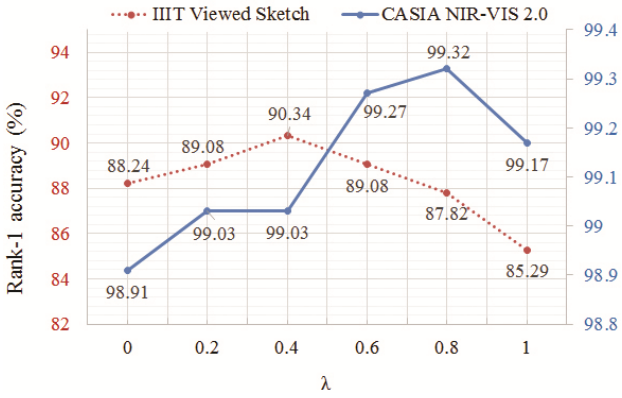

Figure 5: Rank-1 accuracy on IIIT-D Viewed Sketch and CASIA NIR-VIS 2.0 with varied λ.

Comparison to the State of the Art

IIIT-D Viewed Sketch. We show the comparison of RCN and other state-of-the-art methods on the IIIT-D Viewed Sketch in Table 3. VGG (Parkhi, Vedaldi, and Zisserman 2015), Light CNN (Wu et al. 2015), Center Loss (Wen et al. 2016) and CDL (Wu et al. 2017) are CNN-based methods. They only show slightly superior to hand-crafted feature based method MCWLD (Bhatt et al. 2012a). Our RCN outperforms those methods with large margin, i.e. 6.1% higher than MCWLD, 4.99% higher than CDL and 6.27% higher than Center Loss and Light CNN. In addition, RCN also increases the rank-1 accuracy significantly over the pre-trained ResNet-10 and the fine-tuned one.

Forensic Sketch. Table 4 shows the comparison of state-of-the-art methods on the challenging Forensic Sketch dataset. Almost all works (Peng et al. 2017; Klare and Jain 2013; Klare, Li, and Jain 2011; Peng et al. 2016) are based on hand-crafted features instead of CNN features. This may owe to the fact that traditional CNN can not effectively deal

Table 3: Comparison to the state of the art on IIIT-D Viewed Sketch face dataset.

Method	Rank-1 (%)
SIFT (Bhatt et al. 2012a)	76.28
MCWLD (Bhatt et al. 2012a)	84.24
VGG (Parkhi, Vedaldi, and Zisserman 2015)	80.89
Light CNN (Wu et al. 2015)	84.07
Center Loss (Wen et al. 2016)	84.07
CDL (Wu et al. 2017)	85.35
Pre-trained ResNet-10	52.10
Fine-tuned ResNet-10	82.35
RCN-10	**90.34**

with the issues of very large modal discrepancy as well as over-fitting on hundreds of training samples. Indeed, both the pre-trained ResNet-10 and the fine-tuned one show poor performance, which infers that these two issues are not effectively addressed. Our RCN improves the performance by a large margin, i.e. 37.73% over the pre-trained model and 28.30% over fine-tuned one, and get the state-of-the-art performance on the enlarged gallery, which exhibits that RCN potentially alleviates over-fitting and reduces the large modal discrepancy. To the best of our knowledge, it is the first time that CNN feature based method surpasses the others on the Forensic Sketch dataset.

Table 4: Comparison to the state of the art on Forensic Sketch face dataset.

Method	Rank-50 (%)
LFDA (Klare, Li, and Jain 2011)	13.4
D-RS (Klare and Jain 2013)	28.7
G-HFR (Peng et al. 2017)	31.96
SGR-DA (Peng et al. 2016)	54.64
Pre-trained ResNet-10	24.53
Fine-tuned ResNet-10	33.96
RCN-10	**62.26**

CASIA NIR-VIS 2.0. Table 5 shows the comparison of different state-of-the-art methods on CASIA NIR-VIS 2.0 face dataset. From Table 5, we can observe that RCN effectively enhances rank-1 accuracy and VR@FAR=0.1% over pre-trained and fine-tuned ResNet-10. We owe the improvement to the fact that RCN treats the problems of over-fitting and modal discrepancy more effectively. Our RCN also exhibits superior performance to traditional features (Peng et al. 2017; Shi et al. 2017) and other CNN based methods (He et al. 2017; Hu et al. 2018; Wu et al. 2017; 2015; Saxena and Verbeek 2016) on both face recognition and verification tasks. It is worth noting that the our RCN is with similar parameters to CDL, and that RCN increases the rank-1 accuracy of fine-tuned ResNet-10 by 6.54% while CDL gets only 1.47% higher than its backbone model Light CNN-9. RCN is only with 10 convolutional layer and 128-dim output, which demonstrates the superior capacity of our RCN to extract compact and modal invariant features.

Table 5: Comparison to the state of the art on CASIA NIR-VIS 2.0 face dataset.

Method	Rank-1 (%)	VR@FAR=0.1%
Light CNN (Wu et al. 2015)	96.72 ± 0.23	94.77 ± 0.43
Shared, Inter+Intra (Saxena and Verbeek 2016)	85.9 ± 0.9	78
TRIVET (Liu et al. 2016)	95.74 ± 0.52	91.03 ± 1.26
G-HFR (Peng et al. 2017)	85.3 ± 0.03	-
Gabor + HJB (Shi et al. 2017)	91.65 ± 0.89	89.91 ± 0.97
CDL (Wu et al. 2017)	98.62 ± 0.2	98.32 ± 0.05
Synthetic + CNN (Hu et al. 2018)	85.05 ± 0.83	-
WCNN + low-rank (He et al. 2017)	98.7 ± 0.3	98.4 ± 0.4
Pre-trained ResNet-10	86.58 ± 1.36	75.43 ± 1.88
Fine-tuned ResNet-10	97.06 ± 0.43	96.43 ± 0.47
RCN-10	$\mathbf{99.32 \pm 0.15}$	$\mathbf{98.74 \pm 0.24}$

CUHK NIR-VIS. For CUHK NIR-VIS, we set $\lambda = 0.4$ and get the rank-1 accuracy of 99.44%, which is shown in Table 6. The result not only outperforms the pre-trained and fine-tuned ResNet-10 but also is much better than CFDA (Li et al. 2014), MCA (Li et al. 2016) and CEFD (Gong et al. 2017).

Table 6: Comparison to the state of the art on CUHK NIR-VIS face dataset.

Method	Rank-1 (%)
P-RS (Klare and Jain 2013)	75.1
CFDA (Li et al. 2014)	80.19
MCA (Li et al. 2016)	86.43
CEFD (Gong et al. 2017)	83.93
Pre-trained ResNet-10	95.97
Fine-tuned ResNet-10	99.03
RCN-10	**99.44**

Conclusion

In this paper, we introduce an easy-to-implement Residual Compensation Networks (RCN) for heterogeneous face recognition by incorporating a novel residual compensation (RC) module and a modality discrepancy loss (MD loss) into CNN. We fix all convolutional parameters of the backbone CNN and add a light learnable RC module to it to alleviate over-fitting. The RC module also reduces modal discrepancy by adding compensation to NIR/Sketch face features so that its representation can be close to VIS features. The MD loss further reduces the modal discrepancy by minimizing the cosine distance between different modalities. Extensive experiments on IIIT-D Viewed Sketch, Forensic Sketch, CASIA NIR-VIS 2.0 and CUHK NIR-VIS show that our RCN effectively alleviates over-fitting and reduces modal discrepancy, which results in the state-of-the-art performance on these datasets.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (U1613211, 61633021, 61502152) and Shenzhen Research Program (JCYJ20170818164704758, JCYJ20150925163005055, ZDSYS201605101739178).

References

Bhatt, H. S.; Bharadwaj, S.; Singh, R.; and Vatsa, M. 2012a. Memetic approach for matching sketches with digital face images.

Bhatt, H. S.; Bharadwaj, S.; Singh, R.; and Vatsa, M. 2012b. Memetically optimized mcwld for matching sketches with digital face images. *IEEE T-IFS* 7(5):1522–1535.

Cao, K.; Rong, Y.; Li, C.; Tang, X.; and Loy, C. C. 2018. Pose-robust face recognition via deep residual equivariant mapping. *CoRR* abs/1803.00839.

Chen, B. C.; Chen, C. S.; and Hsu, W. H. 2015. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE T-MM* 17(6):804–815.

Gong, D.; Li, Z.; Huang, W.; Li, X.; and Tao, D. 2017. Heterogeneous face recognition: A common encoding feature discriminant approach. *IEEE TIP* 26(5):2079–2089.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 87–102.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, volume 00, 1026–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

He, R.; Wu, X.; Sun, Z.; and Tan, T. 2017. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE T-PAMI* PP(99):1–1.

Hu, G.; Hua, Y.; Yuan, Y.; Zhang, Z.; Lu, Z.; Mukherjee, S. S.; Hospedales, T. M.; Robertson, N. M.; and Yang, Y. 2017. Attribute-enhanced face recognition with neural tensor fusion networks. In *ICCV*, 3764–3773.

Hu, G.; Peng, X.; Yang, Y.; Hospedales, T. M.; and Verbeek, J. 2018. Frankenstein: Learning deep face representations using small data. *IEEE TIP* 27(1):293–303.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding.

Kemelmacher-Shlizerman, I.; Seitz, S. M.; Miller, D.; and Brossard, E. 2016. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 4873–4882.

Klare, B. F., and Jain, A. K. 2013. Heterogeneous face recognition using kernel prototype similarities. *IEEE T-PAMI* 35(6):1410–1422.

Klare, B. F.; Li, Z.; and Jain, A. K. 2011. Matching forensic sketches to mug shot photos. *IEEE T-PAMI* 33(3):639–646.

Li, S. Z.; Yi, D.; Lei, Z.; and Liao, S. 2013. The CASIA NIR-VIS 2.0 face database. In *CVPR Workshops*, 348–353.

Li, Z. F.; Gong, D.; Qiao, Y.; and Tao, D. 2014. Common feature discriminant analysis for matching infrared face images to optical face images. *IEEE TIP* 23(6):2436–2445.

Li, Z.; Gong, D.; Li, Q.; Tao, D.; and Li, X. 2016. Mutual component analysis for heterogeneous face recognition. *ACM T-IST* 7(3):28.

Liao, S.; Yi, D.; Lei, Z.; Qin, R.; and Li, S. Z. 2009. Heterogeneous face recognition from local structures of normalized appearance. In *International Conference on Advances in Biometrics*, 209–218.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*, 3730–3738.

Liu, X.; Song, L.; Wu, X.; and Tan, T. 2016. Transferring deep representation for nir-vis heterogeneous face recognition. In *International Conference on Biometrics (ICB)*, 1–8.

Parkhi, O. M.; Vedaldi, A.; and Zisserman, A. 2015. Deep face recognition. In *BMVC*, 41.1–41.12.

Peng, C.; Gao, X.; Wang, N.; and Li, J. 2016. Sparse graphical representation based discriminant analysis for heterogeneous face recognition. *CoRR* abs/1607.00137.

Peng, C.; Gao, X.; Wang, N.; and Li, J. 2017. Graphical representation for heterogeneous face recognition. *IEEE T-PAMI* 39(2):301–312.

Rebuffi, S. A.; Bilen, H.; and Vedaldi, A. 2018. Efficient parametrization of multi-domain deep neural networks.

Sarfraz, M. S., and Stiefelhagen, R. 2015. Deep perceptual mapping for thermal to visible face recognition. *CoRR*.

Saxena, S., and Verbeek, J. 2016. Heterogeneous face recognition with CNNs. In *ECCV*, 483–491.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Shi, H.; Wang, X.; Yi, D.; Lei, Z.; Zhu, X.; and Li, S. Z. 2017. Cross-modality face recognition via heterogeneous joint bayesian. *IEEE SPL* 24(1):81–85.

Song, L.; Zhang, M.; Wu, X.; and He, R. 2017. Adversarial discriminative heterogeneous face recognition.

Sun, Y.; Wang, X.; and Tang, X. 2014a. Deep learning face representation by joint identification-verification. In *NIPS*.

Sun, Y.; Wang, X.; and Tang, X. 2014b. Deep learning face representation from predicting 10,000 classes. In *CVPR*.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *ECCV*, 499–515.

Wu, X.; He, R.; Sun, Z.; and Tan, T. 2015. A light cnn for deep face representation with noisy labels. *Computer Science*.

Wu, X.; Song, L.; He, R.; and Tan, T. 2017. Coupled deep learning for heterogeneous face recognition. *CoRR*.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Learning face representation from scratch. *Computer Science*.

Zhang, X.; Zhang, L.; Wang, X. J.; and Shum, H. Y. 2012. Finding celebrities in billions of web images. *IEEE T-MM* 14(4):995–1007.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE SPL* 23(10):1499–1503.

Zhang, W.; Wang, X.; and Tang, X. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, 513–520.