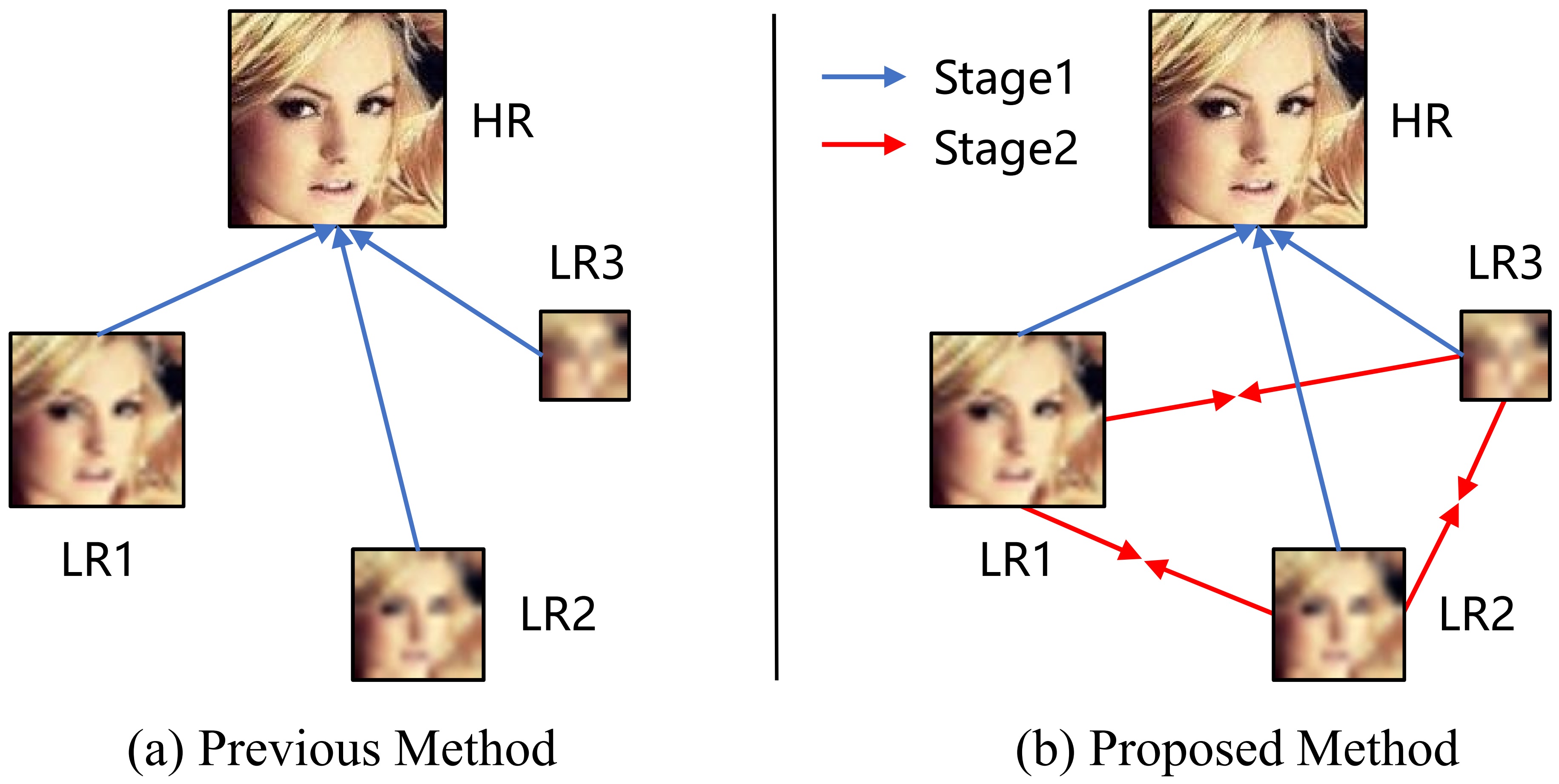
Two-Stage Multi-Scale Resolution-Adaptive Network for Low-Resolution Face Recognition

Paper ID: 1856

Low-resolution face recognition is challenging due to uncertain input resolutions and a lack of distinguishing details in low-resolution (LR) facial images. Resolution-invariant representations must be learned for optimal performance. Existing methods for this task mainly minimize the distance between the representations of the low-resolution (LR) and corresponding high-resolution (HR) image pairs in a common subspace. However, these works only focus on introducing various distance metrics at the final layer and between HR-LR image pairs. They do not fully utilize the intermediate layers or multi-resolution supervision, yielding only modest performance. In this paper, we propose a novel two-stage multi-scale resolution-adaptive network to learn more robust resolution-invariant representations. In the first stage, the structural patterns and the semantic patterns are distilled to provide sufficient supervision. A curriculum learning strategy facilitates the training of LR and HR image matching, smoothly decreasing the resolution of LR images. In the second stage, a multi-resolution contrastive loss is introduced on LR images to enforce intra-class clustering and inter-class separation of the LR representations. By introducing multi-scale supervision and multi-resolution LR representation clustering, our network can produce robust representations despite uncertain input sizes. Experimental results on eight benchmark data sets demonstrate the effectiveness of the proposed method.

# Introduction

Recent years have seen rapid developments in automatic face recognition. The increase in sample numbers and improvement of the loss function has enhanced the performance of high-resolution face recognition on several benchmark data sets such as LFW (Huang et al. 2008), AgeDB-30 (Moschoglou et al. 2017), CFP-FP (Sengupta et al. 2016), IJB-B (Whitelam et al. 2017), and IJB-C (Maze et al. 2018). However, surveillance face recognition is still a complicated challenge because of variations in imaging conditions, particularly resolution. Low-resolution face recognition has two application scenarios: HR-LR and LR-LR. The HR-LR task conducts representation matching between HR galleries and LR probes. In the LR-LR task, both galleries and probes are from LR images.



(a) Previous methods force the representations of LR images to approximate those of HR images. (b) Our proposed method conducts the approximation in stage one and adds the representation clustering of multiple LR images in stage two.

Directly applying HR face recognition methods to LR images results in poor performance due to the lack of facial texture and components. Face recognition methods suitable for HR images cannot extract discriminative representations from LR images. A natural idea is to recover HR images from LR images through the face super-resolution (SR) method (Jiao et al. 2021). SR methods have high computational complexity and large parameter scales, so they are not suitable for real-time deployment. Furthermore, the SR images contain stretching artifacts, which can propagate to face recognition. Other methods (Lu, Jiang, and Kot 2018; Zha and Chao 2019; Yin et al. 2020; Massoli, Amato, and Falchi 2020; Khalid et al. 2020; Fang et al. 2020) map HR images and LR images into a common representation space and minimize the distance between them, as shown in Figure [fig1](a). However, these methods focus on introducing various distance metrics and only minimize the distance between the representations at the final layer for semantic patterns. Intermediate features, which always contain structural patterns and play a key role in performance, are left unleveraged. When generating HR and LR image pairs, the resolution of the LR image is chosen randomly. There are gaps between the domains of LR images with different resolutions, and this random strategy results in slow convergence. Moreover, previous methods primarily minimize the distance between LR and HR image pairs while ignoring the category correlation of multiple LR images. Real-world inputs are usually LR images with uncertain resolution.

In this paper, we propose a two-stage multi-scale resolution-adaptive (TMR) network to fully utilize the cross-resolution supervision of all scales. Specifically, the proposed method includes a multi-scale distillation stage and a multi-resolution clustering stage. In the first stage, a pre-trained HR network is utilized to predict HR features and representations. The distances between the intermediate features of LR and HR images are also minimized. We calculate the multi-scale affinity matrix and maximize mutual information to distill the structural and semantic patterns. Meanwhile, a simple-to-complex curriculum learning strategy facilitates the training of LR and HR image matching. This strategy regards the resolution of samples as the difficulty score and decreases the resolution of LR images smoothly, so convergence occurs more quickly than if resolutions were chosen randomly. While LR representations closely correspond to HR representations at this point, they are not internally clustered by category. In light of this, we propose a novel multi-resolution contrastive (MRC) loss. The goal of MRC loss is to modulate multi-resolution LR representations so those within the same class become more aggregated and those in different classes are farther apart. Consequently, our network predicts resolution-adaptive representations with images of different input sizes. Our method takes the multi-scale feature distillation and multi-resolution representation clustering stages together to provide a new solution to generate less biased and more robust resolution-invariant representations for LR facial recognition.

We conduct LR facial recognition experiments on three realistic LR face data sets, i.e., SCFace, QMUL-SurvFace, and QMUL-TinyFace. We also conduct experiments on down-sampled synthetic LR test sets on five HR face data sets, i.e., LFW, CFP, AgbDB-30, IJB-B, and IJB-C. Experimental results on these benchmark data sets demonstrate the effectiveness of the proposed method. The ablation study also verifies the usefulness of each proposed part. Code will be publicly released after paper acceptance.

Contributions of this work are summarized as follows:

* We propose to minimize the distances between multi-scale intermediate features as well as the final representations of HR and LR images.
* We introduce multi-resolution contrastive (MRC) loss on multi-resolution LR images, enforcing the intra-class clustering and inter-class separation of the LR representations. These two stages form the basis of a two-stage multi-scale resolution-adaptive (TMR) network able to learn robust resolution-invariant representation.
* We conduct experiments on eight widely used benchmark data sets. Results demonstrate the superiority of our method compared to state-of-the-art works.

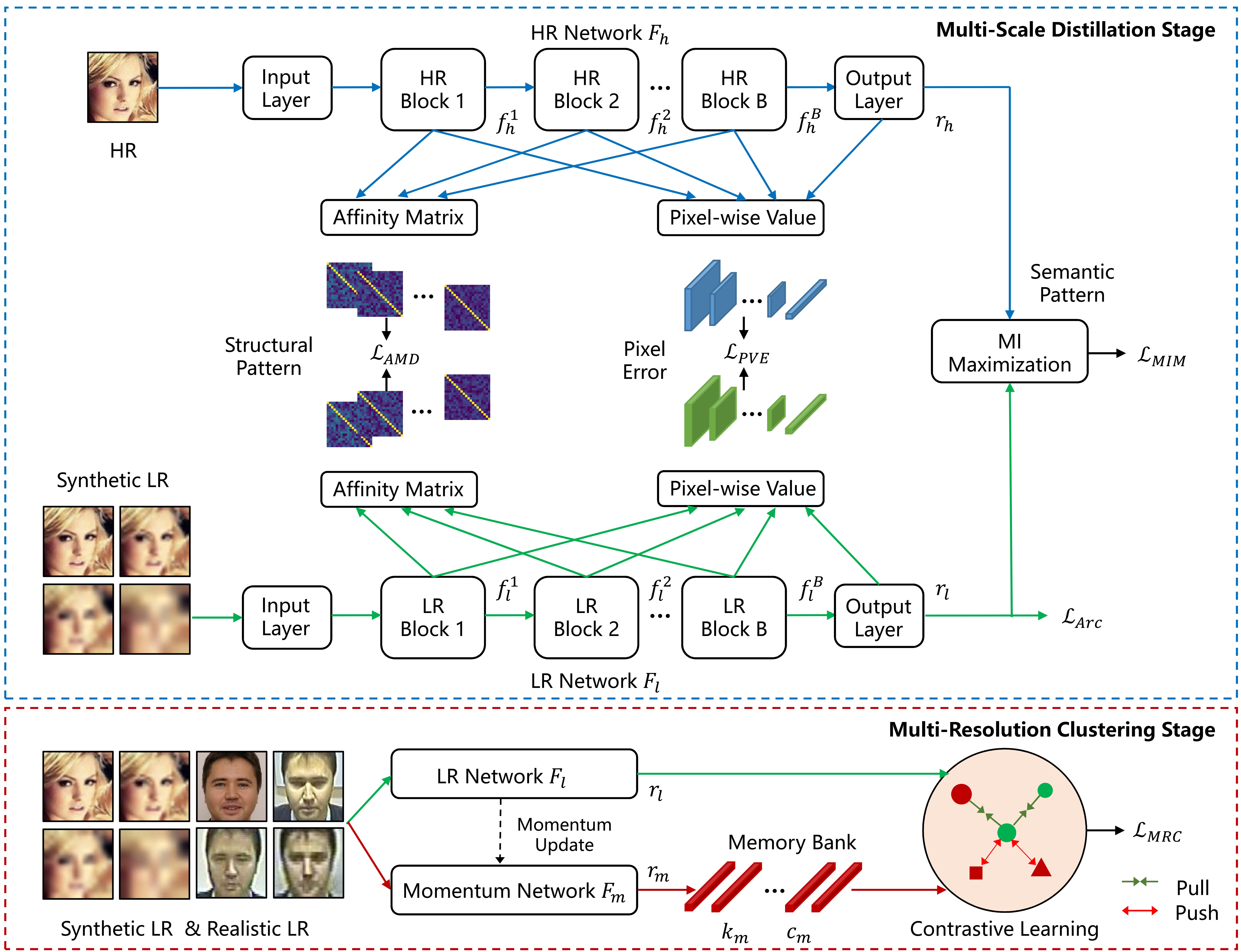


Illustration of our proposed network, which consists of an HR network $F\_h$, an LR network $F\_l$, and a momentum network $F\_m$. In stage one (top), the features $f^i\_h$ and $f^i\_l$ are extracted to align the structural patterns by affinity matrix. The representations $r\_h$ and $r\_l$ are learned to distill the semantic patterns by mutual information maximization (MIM). Pixel value errors are calculated to minimize the distances for both features and representations. In stage two (bottom), the representations $r\_l$ and $r\_m$ are contrasted for the multi-resolution contrastive loss. In the circle on the right, green is from $F\_l$ and red is from $F\_m$. Different shapes represent different classes. Different resolutions of the same sample have the same color and shape but different sizes.

# Related Work

**Low-Resolution Face Recognition.** There are two mainstream methods for low-resolution face recognition. One method applies super-resolution (SR) technology to reconstruct HR images from LR images for HR face recognition. Jiao et al. (Jiao et al. 2021) proposed a dual-domain adaptive translation (DDAT) structure to generate HR images for both synthetic and realistic LR images. DDAT minimizes the domain gap between the synthetic and realistic data sets. SR methods are computationally complex and so are unsuitable for real-time deployment. The synthetic SR images also contain noise, which will cause recognition errors.

The other methods learn similar representations by minimizing the distance between HR and LR representations from the final layer in a common space. Among recent state-of-the-art works of this type, Lu et al. (Lu, Jiang, and Kot 2018) proposed a deep coupled ResNet (DCR) model with center loss to extract robust facial resolution representations. Zha et al. (Zha and Chao 2019) proposed a transferable coupled network (TCN) with triplet loss to learn similar facial representations between HR and LR images. Massoli et al. (Massoli, Amato, and Falchi 2020) introduced the teacher-curriculum (T-C) to approximate representations of the LR images to those of the HR images. In FAN (Yin et al. 2020), a feature adaptation network was proposed to disentangle the representations from HR images and minimize the representation distances between the HR and LR networks. Khalid et al. (Khalid et al. 2020) changed the manner of distance metric by minimizing the KL-divergence between the softmax probabilities of the HR and LR images. Fang et al. (Fang et al. 2020) focused on generating LR faces using a generative adversarial network rather than down-sampling, and then forcing similarity between the representations of the HR and generated LR images.

Besides these two methods, Huang et al. (Huang et al. 2020) proposed a distribution distillation loss (DDL) to narrow the common face recognition performance gap between easy and hard samples. It was also evaluated on LR face recognition test sets. In MIND (Low, Teoh, and Park 2021), a mutual information distillation network (MIND-Net) was proposed to distill the representations between synthetic multi-resolution images and realistic LR images by triplet loss. However, it is difficult to select image pairs according to categories due to the domain gap caused by the absence of the same person. This method performs poorly on both synthetic and realistic test sets.

Existing methods (Lu, Jiang, and Kot 2018; Zha and Chao 2019; Massoli, Amato, and Falchi 2020; Yin et al. 2020; Khalid et al. 2020; Fang et al. 2020) primarily minimize the distances between the representations at the final layer and between HR-LR image pairs. They ignore the intermediate feature constraints and the category correlation of multiple LR representations. Unlike these methods, the first stage of our network distills semantic and structural patterns from the final representations and middle features for the HR-LR task. Furthermore, the second stage considers the contrasting relationship of multi-resolution representations from different classes to modulate representation clustering for the LR-LR task, as shown in Figure [fig1](b).

**Contrastive Learning.** Contrastive learning performs well and is frequently used for representation learning. Chen et al. (Chen et al. 2020) proposed SimCLR, a simple framework that introduced contrastive learning into representation learning. He et al. (He et al. 2020) proposed a momentum contrast (MoCo) mechanism to build significant and consistent dictionaries for unsupervised learning. Previous mainstream contrastive loss methods typically consider the contrast from two views. However, we aim to learn the representations of images with varying resolutions. Inspired by MoCo, we propose a novel multi-resolution contrastive (MRC) loss for robust clustering of multiple LR representations. Our MRC loss extends two views to multiple views.

# Proposed Method

This section presents the problem statement and the facial representation learning network. Then, the first stage of multi-scale distillation and the second stage of multi-resolution clustering are introduced. The overall network is illustrated in Figure [fig2].

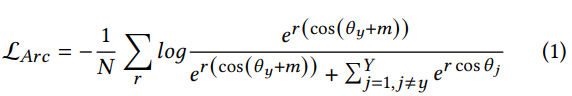
## Problem Statement

The training set consists of $N$ original HR images. The HR images are down-sampled to synthesize four LR images. The training set $\mathcal{D}\_{s}$ is composed of the original HR and synthetic LR images $\big\{ I\_h, I^{1}\_{sl}, I^{2}\_{sl}, I^{3}\_{sl}, I^{4}\_{sl},y \big\}$ , where $ I\_h \in \mathbb{R}^{H\times W \times 3}$ represents the original HR image, $I^{i}\_{sl} \in\mathbb{R}^{H^i\times W^i\times 3}$ represents the $i$-th synthetic LR image, and $y \in \left\{ 1,2,...,Y \right\}$ represents the $y$-th class. There is another training set $\mathcal{D}\_{r}$ with $M$ realistic LR images $\left\{ I\_{rl},y\_{rl} \right\} $, where $I\_{rl}$ represents the realistic LR image with uncertain resolution. During testing, the images of uncertain resolution are inputted to output their representations for face recognition.

## Facial Representation Learning Network

As illustrated in the top of Figure \ref{fig2}, the facial representation learning network mainly consists of two neural networks with the same structure. Each network is divided into an input layer, $B$ stacked blocks, and an output layer. The first $ F\_{h} $ is the HR network, into which the HR image is input to obtain multi-scale HR features $\big\{ f\_h^i \big\}\_{i=1}^B$ and the output HR representation $r\_h = F\_h(I\_h)$, where $ f\_h^i$ denotes the features at the $i$-th block. The second network $ F\_{l} $ is the LR network, which uses the LR image $I\_{l} \in \left\{ I\_{sl},I\_{rl}\right\} $ as the input to obtain multi-scale LR features $\big\{ f\_l^i \big\}^B\_{i=1}$ and the output LR representation $r\_l = F\_l(I\_l)$. Before outputting, the representations $r\_h$ and $r\_l$ are modified by $L2$ normalization.

The HR network $ F\_{h} $ is pre-trained on an HR image data set to learn discriminative representations. The additive angular margin loss $\mathcal{L}\_{Arc}$ is adapted from ArcFace \cite{deng2019arcface} and can be presented as:



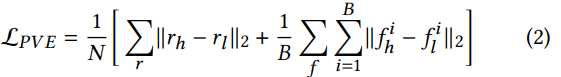
in which $r$ denotes the representation belonging to the $y$-th class. $W$ is the last fully connected layer and $W\_j$ is the $j$-th column of the $W$. $\theta\_{j}$ is the angle between the weight $W\_j$ and the representation $r$. $C$ and $m$ are the class number and the additive angular margin penalty, respectively.

After pre-training, the parameters of the HR network $ F\_{h} $ are fixed. The LR network $ F\_{l} $ is initialized with the pre-trained parameters from $ F\_{h} $. This allows the LR network to extract representations from HR images. The LR network is then trained in two stages, a multi-scale distillation stage and a multi-resolution clustering stage.

## Multi-Scale Distillation Stage

We propose a multi-scale distillation stage to fully utilize the multi-scale supervision from the HR network. Unlike existing methods, this stage aligns the features at multiple scales as well as the representations of HR and LR images. This alignment is implemented through three types of distillations: pixel-wise value error (PVE), affinity matrix distillation (AMD), and mutual information maximization (MIM). There are various distance metrics between the HR and LR representations; we use PVE to minimize the pixel-wise distance between the final representations. Our method also applies PVE to the intermediate features, and further applies AMD to the paired HR and LR features of intermediate layers to align their structural patterns. Mutual information maximization (MIM) increases the approximation of the representation distribution by reducing the distance between the joint and marginal distributions of the HR and LR representations. The training sequence is provided under the curriculum learning strategy (CLS). During this stage, the LR network is trained with the HR images and synthetic LR images from $\mathcal{D}\_{s}$.

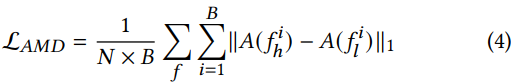
**Pixel-Wise Value Error.** To distill the local information in pixel-wise values from HR images to LR images, we minimize the Euclidean distance $\mathcal{L}\_{Pix}$ between paired multi-scale features $\big\{ f\_h^i ,f\_l^i \big\}^B\_{i=1}$ and representations $\left\{ r\_h,r\_l \right\} as follows:



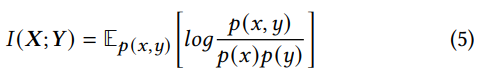
**Affinity Matrix Distillation.** As the resolution decreases, the texture and components of the face (eyes, eyebrows, mouth, etc.) become blurrier. However, the shape and contour information remains unchanged and the structures of LR images are consistent with those of HR images. We were motivated by \cite{wang2018non}, who posit that non-local operations like the affinity matrix can enlarge the receptive field and calculate the relationship between pixels, thus reflecting the global correlation of features pixel to pixel and the resolution-invariant structural pattern. Specifically, given a feature $f \in\mathbb{R}^{h\times w\times c}$, reshape operation $R(\cdot)$ pulls it into a two-dimensional vector $R(f) \in\mathbb{R}^{hw\times c}$. Then, the affinity matrix $A(f)$ is defined as:



where $\sigma(\cdot)$ is a softmax function, $\otimes$ is the matrix multiplication, and $T$ is the matrix transpose operation. To align the structural pattern of the HR and LR features, we minimize the Manhattan distance $\mathcal{L}\_{AMD}$ between their affinity matrices:

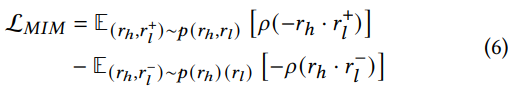


**Mutual Information Maximization.** Mutual information (MI) is an elementary measurement to quantify the dependence of two random variables. The MI between random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as:



where $p(x,y)$ is the joint probability distribution and $p(x)p(y)$ are their marginal distributions.

Maximizing mutual information between different views strengthens the dependence and reduces the discrepancy of multi-view variable distribution to learn view-robust representations. In this work, we maximize the cross-resolution MI to facilitate the semantic pattern consistency of the HR and LR representations. Since it is challenging to calculate MI directly, most works estimate the lower bound of mutual information. Inspired by DIM \cite{hjelm2018learning}, we use a Jensen-Shannon MI estimator, which maintains a good balance between computational complexity and performance. Within a mini-batch, given HR representation $r\_h$, its homogeneity positive match $r^+\_l$, and its heterogeneity negative match $r^-\_l$, the positive pair $\big\{ r\_h,r^+\_l \big\}$ represents a sample from the joint distribution $p(r\_h,r\_l)$ and the negative pair $\big\{r\_h,r^-\_l\big\}$ represents a sample from the marginal distribution $p(r\_h)(r\_l)$. MI estimation transfers semantic patterns by considering the category relationship to represent the two distributions. Maximizing the lower bound of mutual information is equivalent to minimizing the loss of $\mathcal{L}\_{MIM}$:



where $\rho(z)=log(1+e^z)$ is the softplus function.

**Loss Function of Stage One.** For paired HR and LR images $\big\{ I\_h,I^i\_{sl} \big\}$ , the final loss in the multi-scale distillation stage is:



where $\varphi\_{1}$, $\varphi\_{2}$, and $\varphi\_{1}$ are three hyperparameters for weighing different losses.

**Curriculum Learning Strategy.** Current methods randomly choose the resolution of the LR image when generating HR and LR image pairs. However, face images with lower resolution have fewer details and thus have larger domain shifts from the HR images. More significant shifts in lower resolution images will interfere with the representation distribution learning of the higher resolution images. The domain shifts vary when LR image resolution is randomly chosen, slowing down the convergence during training. Instead of choosing LR images with a random resolution, a simple-to-complex curriculum learning strategy (CLS) is designed to facilitate the training of HR and LR image matching. This strategy regards the resolution of samples as the difficulty score and decreases the resolution of LR images smoothly, so convergence occurs more quickly.

## For the starting epochs, the highest resolution LR images are paired with HR ones. Step by step, LR images with lower resolutions are added to the choices. Specifically, in the $i$-th step, our networks are trained with HR images $I\_h$ and LR images in $\big\{I^{1}\_{sl}, ..., I^{i}\_{sl}\big\}$ for $e^i$ epochs, where $\big\{I^{i}\_{sl} \big\}$ is the LR image with $i$-th low resolution. Therefore, the loss function in stage one becomes $\sum\_{j=1}^i \mathcal{L}\_{Stage1}(I\_h,I^j\_{sl})$. The higher LR samples help reduce large domain shifts by gradually making the lowest resolution representations close to HR representations.

## Multi-Resolution Clustering Stage

After stage one, the representations of multiple LR images from the same sample are clustered around the HR representations. However, the domain discrepancy among different LR representations remains. To tackle this obstacle, we propose a multi-resolution representation clustering stage. During this stage, the LR network is first trained with synthetic LR images from data set $\mathcal{D}\_{s}$, and then fine-tuned on realistic LR data set $\mathcal{D}\_{r}$.

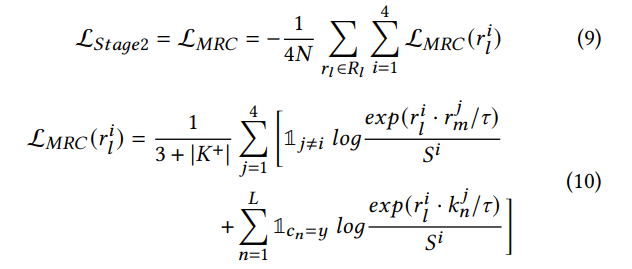
**Training with Synthetic Low-Resolution Images.** Recently, contrastive learning has achieved excellent performance in many self-supervised and unsupervised works by learning representations with high intra-class compactness and inter-class discrepancy. The first stage of our method distills facial information from HR to LR images while neglecting the category correlations among multiple LR images. In light of this, we propose a contrastive learning framework with multi-resolution contrastive (MRC) loss. Unlike the original contrastive loss, MRC loss extends the regular two views into multiple views. The intra-class compactness is maximized to cluster the representations of multiple LR samples from the same person and the same class. Inter-class discrepancy is maximized to push the representations of heterogeneous samples farther away from the decision boundary.

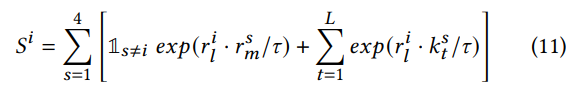
Specifically, the LR network is replicated as the momentum network $F\_m$. For the training set $\mathcal{D}\_{s}$, a key memory bank $K$ for LR representations of four resolutions and a category memory bank $C$ for the corresponding labels are maintained,



where $L$ is the length of the memory bank. Both memory banks are dynamic queues, initialized with four LR representations and labels sampled randomly from the data set $\mathcal{D}\_{s}$.

A set of LR samples $\big\{ I^1\_{sl}, I^2\_{sl}, I^3\_{sl}, I^4\_{sl}, y \big\}$ is input into the LR network and momentum network to obtain the representations $\big\{ r^1\_{l}, r^2\_{l},r^3\_{l}, r^4\_{l} \big\}$ and $\big\{ r^1\_{m}, r^2\_{m}, r^3\_{m}, r^4\_{m} \big\}$. Assuming $r^i\_{l}$ is an anchor, its positive pairs are the representations $\big\{ r^j\_{m} | j \neq i \big\}$ of the same sample at different resolutions and the representations $K^+=\big\{ k^j\_n | c\_n= y,j=1,2,3,4 \big\}$ of the same class at four resolutions from memory bank $K$. Its negative pairs are representations $K^-=\big\{ k^j\_n | c\_n\neq y,j=1,2,3,4 \big\}$ of the different classes at four resolutions, also from memory bank $K$. The $\mathcal{L}\_{MRC}$ loss is defined as:





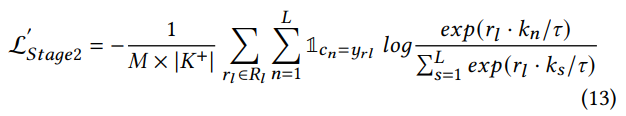
where $i$ and $j$ are the resolution superscripts of the anchor and contrastive samples, respectively. $R\_l$ denotes the LR representation set, which contains four LR representations. $\tau >0$ is a temperature parameter.

After a mini-batch is factored into the calculation, the generated representations of four resolutions $\left\{ r^1\_{m}, r^2\_{m}, r^3\_{m}, r^4\_{m} \right\}$ and their labels are enqueued to the key memory bank $K$ and category memory bank $C$. The momentum updates through the LR network as follows:



where $\theta\_m$ and $\theta\_l$ are the parameters of $F\_m$ and $F\_l$ respectively. $\lambda \in \left[0,1 \right) $ is a momentum coefficient.

**Fine-Tuning on Realistic Low-Resolution Images.** For realistic LR sample $\left\{ I\_{rl},y\_{rl} \right\} \in \mathcal{D}\_{r}$, our key memory bank $K=\left[k\_{n} \right]^L\_{n=1}$ and category memory bank $C=\left[c\_{n}\right]^L\_{n=1}$ also maintain the representations and labels. Since the true resolution of these samples is unknown, different images of the same person are treated as multi-resolution. Given an anchor $r\_l$, its positive pairs ${K}^+ = \left\{ k\_{n} | c\_n= y\_{rl} \right\}$, and its negative pairs ${K}^- = \left\{ k\_{n} | c\_n\neq y\_{rl} \right\}$, a regular contrastive loss function is used:



where $R\_l$ denotes the realistic LR representation set.

To summarize, during the whole training phase, the network is trained with the HR and LR images from $\mathcal{D}\_{s}$ by $\mathcal{L}\_{Stage1}$. Then it is trained with multiple LR images, also from $\mathcal{D}\_{s}$ by $\mathcal{L}\_{Stage2}$. Finally, it is fine-tuned on each $\mathcal{D}\_{r}$ by $\mathcal{L}^{'}\_{Stage2}$.

# Experiments

**Data sets.** Existing works train the coupled HR and LR networks with different data sets. Following ArcFace\cite{deng2019arcface}, our network is trained on the MS1M-ArcFace. Original HR images are resized from 112$\times$112 to 56$\times$56, 28$\times$28, 14$\times$14, and 7$\times$7 to synthesize LR images via bicubic interpolation and rescale them to the initial input size. Experiments are conducted on eight benchmark data sets, i.e. SCFace \cite{grgic2011scface}, QMUL-SurvFace \cite{cheng2018surveillance},QMUL-TinyFace \cite{cheng2018low}, LFW \cite{huang2008labeled}, CFP \cite{sengupta2016frontal}, AgbDB-30 \cite{moschoglou2017agedb}, IJB-B \cite{whitelam2017iarpa}, and IJB-C \cite{maze2018iarpa}.

SCFace, QMUL-SurvFace, and QMUL-TinyFace are realistic face data sets. We fine-tune their training sets and use their testing sets to compare our results with state-of-the-art methods. The SCFace data set is captured by five video surveillance cameras located at three different distances: 4.20 m (d1), 2.60 m (d2), and 1.00 m (d3). It includes 4,160 images of 130 subjects. Following previous works \cite{lu2018deep}, we randomly sample 80 subjects for fine-tuning and use the rest for testing Rank-1 accuracy. The QMUL-SurvFace data set includes 463,507 images of 15,573 distinct identities. Following Cheng et al.'s work \cite{cheng2018surveillance}, we evaluate the metrics of TPIR20(\%)@FPIR and AUC for face identification. This yields 60,294 images from 5,319 subjects as the gallery; 60,423 images from 5,319 subjects as mated probes; and 12,1736 distractor images as unmated probes. We also evaluate the metrics of TAR@FAR, AUC, and Mean.Acc for face verification. There are 5,320 positive pairs and 5,320 negative pairs. The QMUL-TinyFace data set includes 169,403 images of 5,139 identities. Following Cheng et al.'s work \cite{cheng2018low}, we evaluate the metrics of Rank-1, 20, 50, and mAP for face identification. There are 4,443 images from 2,569 subjects as the gallery; 3,728 images from 2,569 subjects as mated probes; and 153,428 distractor images as unmated probes.

The LFW, CFP, AgbDB-30, IJB-B, and IJB-C are HR face testing sets. We resize them to the same four low resolutions as the training set for ablation studies. The LFW data set contains 13,233 web-collected face images from 5,749 different subjects. The CFP data set consists of 10 frontal and four profile images of 500 individuals for 7,000 images. The AgbDB-30 data set contains 16,488 images of 568 distinct subjects. Rank-1 accuracy is evaluated on these three test sets. The IJB-B data set is composed of 1,845 identities with 21,800 still images and 55,000 frames. For face verification, 10,270 positive pairs and 8M negative pairs are matched. The IJB-C data set extends IJB-B, including 31,334 images of about 3,500 subjects and 117,542 unconstrained video frames. For face verification, 19,557 positive pairs and 15,638,932 negative pairs are matched. Following ArcFace\cite{deng2019arcface}, we evaluate TAR (@FAR=1e-4) on these two test sets. Furthermore, according to the DCR \cite{lu2018deep}, the sizes of the images from LFW are also resized to 8$\times$8, 12$\times$12, 16$\times$16, 20$\times$20, and 112$\times$96 to compare with existing works.

**Implementation Details.** We adopt SE-LResNet50E-IR \cite{deng2019arcface} as the backbone of the HR and LR networks. When pre-training the HR network, the input size is 112$\times$112 and the output is a $512-d$ representation. Then the parameters of the well-trained HR network are copied to the LR network. In stage one, stochastic gradient descent (SGD) is taken as the optimizer, with the initial learning rate $1 \times {10}^{-3} $, momentum 0.9, and weight decay $5 \times {10}^{-4} $. We proceed to train with 16 epochs and divide the learning rate by 10 on \{4, 8, 12\}-$th$ epochs for better convergence. Each $e^i$ is set to four. In stage two, we also use the SGD optimizer with the learning rate of $1 \times {10}^{-5} $ and four epochs.

Following \cite{fang2020generate}, the baseline is the LR network trained by ArcFace loss with the synthetic LR images. The optimal values for the hyper-parameters, i.e., {$\varphi\_{1}$, $\varphi\_{2}$, $\varphi\_{3}$, $m$, $\tau$, and $\lambda$}, are searched and found to be \{0.1, 100, 10, 0.5, 0.1, 0.99\}. All experiments are conducted with PyTorch 1.6.0 on two NVIDIA GeForce RTX 3090 GPU with 48GiB memory.

## Comparisons to State-of-the-Art Methods

We compared our method with existing top performers on four benchmark data sets: LFW, SCFace, QMUL-SurvFace, and QMUL-TinyFace. On each data set, our method is only compared to methods with reported results on this data set. Note that on the three realistic LR data sets, the TMR method with only the first stage is fine-tuned by ArcFace loss for a fair comparison with the TMR which has both stages and is fine-tuned by $\mathcal{L}^{'}\_{Stage2}$

**Results on LFW.** The comparison between the experimental results of the proposed approach and other existing methods on the LFW data set is presented in Table [table1]. As shown in the table, the proposed method outperforms prior works across all input sizes. In particular, our TMR network achieves the highest accuracy of 95.9%, 97.9%, 98.7%, 99.1%, and 99.7%, as well as an improvement of 1.3%, 0.8%, 0.4%, 0.6%, and 0.4% over Baseline. The improvement is more significant at smaller resolutions, indicating that our model is adaptable at representation learning with low resolutions. Both DCR and TCM minimize the distance between the representations at the final layer only. The intermediate features, which always contain the structural patterns and act as significant differentiating factors in face recognition, are left unconstrained. Our TMR network introduces supervision on multi-scale intermediate features to fully capture the guidance from HR to LR images, thus obtaining more robust representations. The results of experiments on the LFW data set show that our approach is especially beneficial at lower resolutions.

Face verification accuracy (%) of different methods using different probe sizes on the LFW data set. Our baseline is the LR network trained by ArcFace loss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods | 88 | 1212 | 1616 | 2020 | 11296 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ResNet (He et al. 2016) | 72.7 | 84.1 | 92.3 | 95.4 | 98.7 |
| ResNet-FT | 88.9 | 93.8 | 95.9 | 96.8 | 98.8 |
| Trunk (Lu, Jiang, and Kot 2018) | 92.2 | 93.6 | 95.5 | 96.8 | 98.4 |
| DCR (Lu, Jiang, and Kot 2018) | 93.6 | 95.3 | 96.6 | 97.3 | 98.7 |
| TCN (Zha and Chao 2019) | 90.5 | 94.7 | 97.2 | 97.8 | n/a |
| Baseline | 94.6 | 97.1 | 98.3 | 98.5 | 99.3 |
| Ours(Stage1) | 95.5 | 97.5 | 98.5 | 99.0 | 99.6 |
| Ours(Stage1+Stage2) | **95.9** | **97.9** | **98.7** | **99.1** | **99.7** |

Rank-1 IR(%) of face identification on SCFace data set. Testing without fine tuning is indicated by ’w/o FT.’

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | d1 | d2 | d3 | avg. |
| DCR (Lu, Jiang, and Kot 2018) | 73.30 | 93.50 | 98.00 | 88.27 |
| TCN (Zha and Chao 2019) | 74.60 | 94.90 | 98.60 | 89.37 |
| C-T (Massoli, Amato, and Falchi 2020) | 45.10 | 85.90 | 96.10 | 75.70 |
| FAN (Yin et al. 2020) | 77.50 | 95.00 | 98.30 | 90.30 |
| FAN w/o FT | 62.00 | 90.00 | 94.80 | 82.30 |
| DDL (Huang et al. 2020) | 86.80 | 98.30 | 98.30 | 94.40 |
| RAN (Fang et al. 2020) | 81.30 | 97.80 | 98.80 | 92.63 |
| RAN w/o FT | 70.50 | 96.00 | 98.00 | 88.17 |
| MIND (Low, Teoh, and Park 2021) | 81.75 | 98.00 | 99.25 | 93.00 |
| DA (Khalid et al. 2020) | 88.30 | 98.30 | 98.60 | 95.00 |
| Baseline w/o FT | 76.75 | 95.00 | 96.25 | 89.33 |
| Ours(Stage1) w/o FT | 78.25 | 96.50 | 97.25 | 90.67 |
| Ours(Stage1+Stage2) w/o FT | 79.25 | 97.00 | 97.75 | 91.33 |
| Baseline | 85.50 | 96.25 | 97.50 | 93.08 |
| Ours(Stage1) | 89.00 | 98.25 | 99.00 | 95.42 |
| Ours(Stage1+Stage2) | 91.25 | **99.50** | **99.50** | **96.75** |

**Results on SCFace Data set.** Our method is compared to state-of-the-art methods such as DCR (Lu, Jiang, and Kot 2018), TCN (Zha and Chao 2019), T-C (Massoli, Amato, and Falchi 2020), FAN (Yin et al. 2020), DDL (Huang et al. 2020), RAN (Fang et al. 2020), MIND (Low, Teoh, and Park 2021), and DA (Khalid et al. 2020). The Rank-1 accuracy on the SCFace data set is shown in Table [table2]. Our proposed TMR network outperforms others by 0.85%, 1.00%, 0.25%, and 1.28% in d1, d2, and d3 distances and the average. Furthermore, our method improves over Baseline by 5.75%, 3.25%, 2.00%, and 3.67%. The version without fine tuning also outperforms RAN without fine tuning by 8.75%, 1.00%, and 3.16% in d1 and d2 distance and average. This is likely because previous works focus on representation matching between HR and LR images alone, ignoring the correlation between multiple LR images. In contrast, we propose a novel multi-resolution contrastive loss to force multi-resolution LR representations of the same class to cluster and representations with different classes to disperse. Consequently, our network predicts resolution-adaptive representations and achieves better performance. The performance reflects the superiority of our TMR network in the open-set scenario.

Face identification results on QMUL-SurvFace.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 30% | 20% | 10% | 1% |  |
| VggFace (Parkhi, Vedaldi, and Zisserman 2015) | 6.5 | 4.8 | 2.5 | 0.2 | 9.6 |
| FaceNet (Schroff, Kalenichenko, and Philbin 2015) | 12.7 | 8.1 | 4.3 | 1.0 | 19.8 |
| DeepID2 (Yi, Wang, and Tang 2014) | 12.8 | 8.1 | 3.4 | 0.8 | 20.8 |
| SphereFace (Liu et al. 2017) | 21.3 | 15.7 | 8.3 | 1.0 | 28.1 |
| CentreFace (Wen et al. 2016) | 27.3 | 21.0 | 13.8 | 3.1 | **37.3** |
| RAN (Fang et al. 2020) | 26.5 | 21.6 | 14.9 | 3.8 | 32.3 |
| Baseline | 20.8 | 16 | 10.1 | 2.2 | 29.3 |
| Ours(Stage1) | 25.2 | 21.0 | 15.7 | 6.2 | 33.4 |
| Ours(Stage1+Stage2) | **27.4** | **23.2** | **17.8** | **7.7** | 35.6 |

Face verification results on QMUL-SurvFace.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 30% | 10% | 1% | 0.1% |  | Acc(%) |
| VggFace (Parkhi, Vedaldi, and Zisserman 2015) | 83.2 | 63.0 | 20.1 | 4.0 | 85.0 | 78.0 |
| FaceNet (Schroff, Kalenichenko, and Philbin 2015) | 94.6 | 79.9 | 40.3 | 12.7 | 93.5 | 85.3 |
| DeepID2 (Yi, Wang, and Tang 2014) | 80.6 | 60.0 | 28.2 | 13.4 | 84.1 | 76.1 |
| SphereFace (Liu et al. 2017) | 80.0 | 63.6 | 34.1 | 15.6 | 83.9 | 77.6 |
| CentreFace (Wen et al. 2016) | 95.2 | 86.0 | 53.3 | **26.8** | 94.8 | 88.0 |
| FAN (Yin et al. 2020) | 71.3 | 44.6 | 12.9 | 2.8 | 76.9 | 70.9 |
| DDAT (Jiao et al. 2021) | 90.4 | 75.5 | 40.4 | 16.4 | n/a | 83.6 |
| Baseline | 90.1 | 69.7 | 27.5 | 7.7 | 89.8 | 81.5 |
| Ours(Stage1) | 94.0 | 82.1 | 45.6 | 16.8 | 93.5 | 86.2 |
| Ours(Stage1+Stage2) | **95.8** | **88.0** | **57.7** | 20.7 | **95.4** | **89.1** |

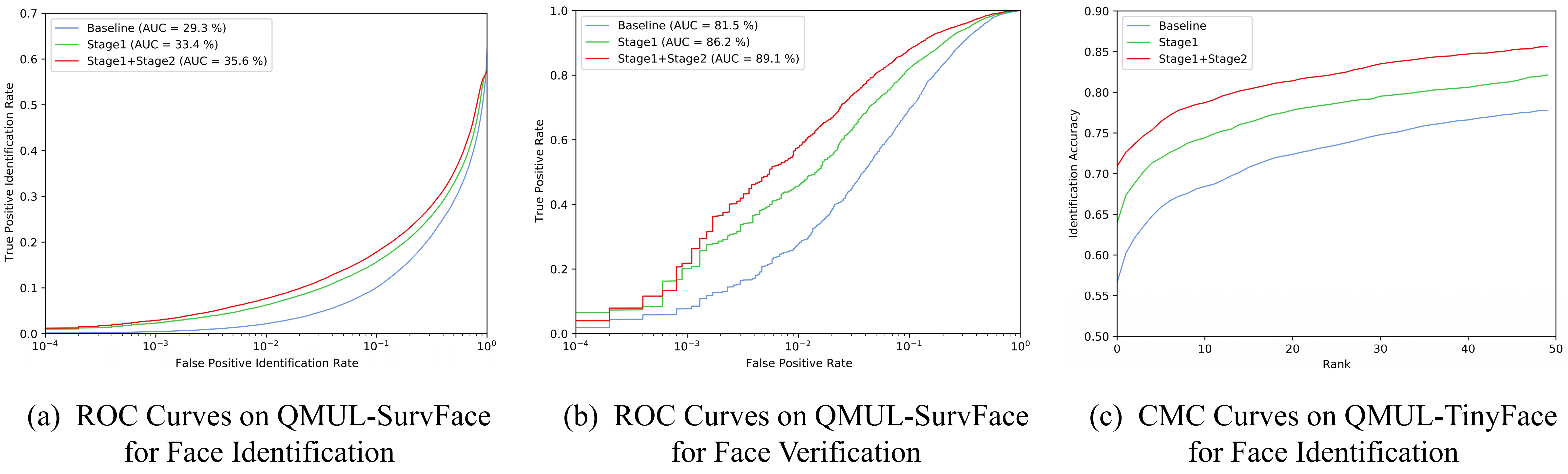
**Results on QMUL-SurvFace and QMUL-TinyFace Data Sets.** On the QMUL-SurvFace data set, we evaluate the metrics of TPIR20(%)@FPIR and AUC for face identification and TAR@FAR, AUC, and Mean.Acc for face verification. As shown in Table [table3], our method outperforms the results of HR methods directly cited from QMUL-SurvFace by 0.1%, 2.2%, 4%, and 4.6% on the TPIR during face identification testing. TMR also achieves better performance than HR methods during face verification, as shown in Table [table4]. These results indicate that it is challenging to directly extract discriminative representations from LR images using HR face recognition methods due to the details missing from LR images. Instead, our approach exploits the guidance from HR images to LR images in the common space. Although RAN uses deep networks rather than down-sampling to generate LR inputs of different sizes, our method improves face identification performance by 3% on the AUC in Table [table3]. This validates that the method of fine-tuning the clustering distribution with realistic LR images is better than using generated LR images. Moreover, Table [table4] shows that our method achieves comparable face verification results compared to FAN and DDAT. In summary, our multi-resolution clustering method is more suitable for identity-robust representation learning than the disentangled-based and adversarial-based methods.

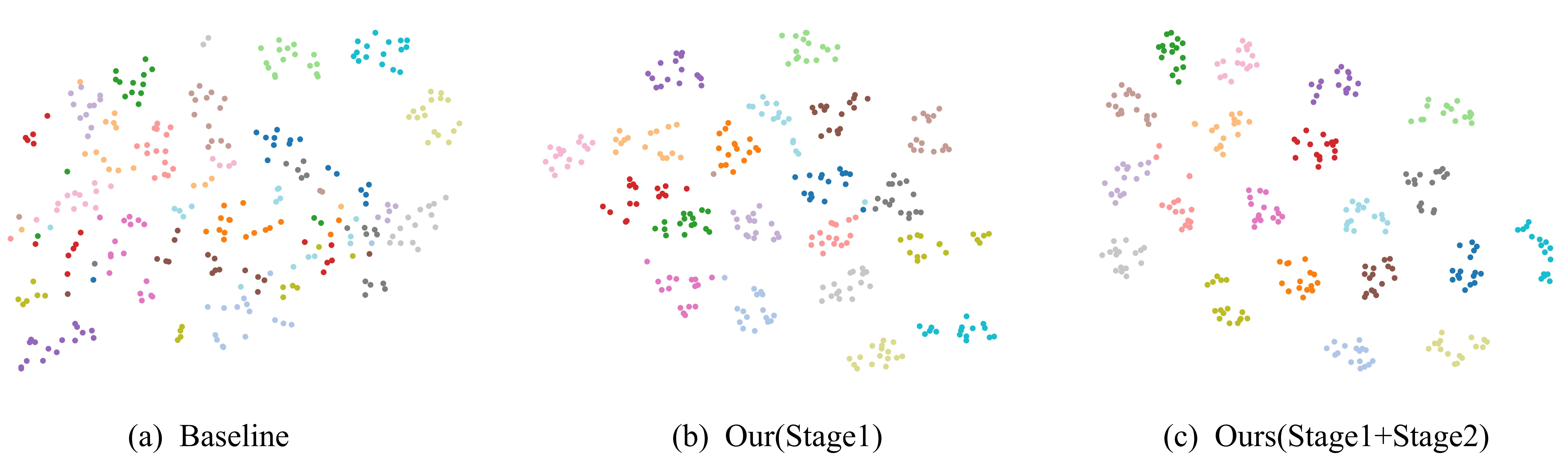
On the QMUL-TinyFace data set, we evaluate the metrics of Rank-1, 20, 50, and mAP for face identification. As shown in Table [table5], our approach achieves the best Rank-1 accuracy of 70.9%, Rank-50 accuracy of 85.6%, and mAP of 64.8% for face verification. Although the DA method trained on synthetically down-sampled LR images performs better than our method in Rank-20 accuracy, our method achieves better performance in all the other metrics.

Since most faces are barely visible in the QMUL-SurvFace and QMUL-TinyFace data sets, facial recognition on these two data sets is rather challenging. Despite that, our method yields significant gains over the Baseline, demonstrating the effectiveness of our method in real world uncooperative surveillance scenes.

Face identification results on QMUL-TinyFace.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methods | Rank-1 | Rank-20 | Rank-50 | mAp |
| VggFace (Parkhi, Vedaldi, and Zisserman 2015) | 30.4 | 44.5 | 48.4 | 24.6 |
| DeepID2 (Yi, Wang, and Tang 2014) | 17.4 | 25.2 | 28.3 | 12.1 |
| SphereFace (Liu et al. 2017) | 22.3 | 35.5 | 40.5 | 16.2 |
| CentreFace (Wen et al. 2016) | 32.1 | 44.5 | 48.4 | 24.6 |
| CSRI (Cheng, Zhu, and Gong 2018a) | 44.8 | 60.4 | 65.1 | 36.2 |
| C-T (Massoli, Amato, and Falchi 2020) | 58.6 | 73.0 | 76.3 | 52.7 |
| MIND (Low, Teoh, and Park 2021) | 66.8 | n/a | n/a | n/a |
| DA (Khalid et al. 2020) | 70.4 | **82.2** | 85.4 | 63.2 |
| Baseline | 56.6 | 72.2 | 77.8 | 53.1 |
| Ours(Stage1) | 64.0 | 77.5 | 82.1 | 60.4 |
| Ours(Stage1+Stage2) | **70.9** | 81.3 | **85.6** | **64.8** |





## Ablation Study and Analysis

As demonstrated in Section 3, our framework consists of two stages. Ablation studies are conducted to validate the effectiveness of each proposed part. Note that when testing on three realistic data sets, the LR network is fine-tuned on them in the second stage, which affects the performance. Therefore, for the ablation studies of the proposed modules such as affinity matrix distillation, experiments in Table [table6] are performed only on the five HR image sets. The HR images are resized to four resolutions and experiments are conducted on the generated synthetic LR images. Here, we present detailed analyses of the results in Table [table1]-[table6].

**Effect of Affinity Matrix Distillation and Mutual Information Maximization.** An affinity matrix distillation (AMD) is introduced in the first stage to utilize the guidance from the HR to LR images at the intermediate features. As shown in Table [table6], the AMD has some benefits over the Baseline, indicating that the structural patterns of facial contour and shape have a significant influence on resolution-invariant feature learning. We also maximize the mutual information as an alignment manner between the HR and LR representation. From Table [table6], the mutual information maximization (MIM) yields improvements over the Baseline. This is because MIM increases the approximation of the representation distribution by reducing the distance between the joint and marginal distributions from HR and LR representations. We consider category labels to estimate these two distributions for the distillation of semantic patterns.

**Effect of Curriculum Learning Strategy.** Existing methods randomly chose the resolution of the LR image when generating LR and HR image pairs, impeding convergence. In our method, a simple-to-complex curriculum learning strategy (CLS) is designed to facilitate the training of LR and HR image matching. From Table [table6], the curriculum learning strategy reflects a particular effect, especially at lower resolution. However, at the 5656 size of IJB-B and IJB-C testing, using the curriculum learning strategy yields worse results than Baseline. This shows that the simple-to-complex strategy sacrifices some higher resolution effects to smoothly adapt to the lower resolution.

**Effect of Multi-Resolution Contrastive Loss.** Although the representations of multiple LR images are clustered around the HR representations after stage one, the domain discrepancy among different LR representations remains. We introduce a multi-resolution contrastive loss to address this problem. Tables 1-6 prove that our network performs better with both stages than it does with stage one alone. Figure [fig3] presents the ROC curves on the QMUL-SurvFace data set for face identification and verification, and the CMC curves on the QMUL-TinyFace data set for face identification. These results verify that MRC loss in the second stage facilitates the LR network by learning representations with high intra-class compactness and inter-class discrepancy.

**Visualization.** In addition to quantitative results, Figure [fig4] provides the t-SNE visualization of the learned representations for the Baseline and our methods at different stages. Twenty subjects were randomly selected from the test set of SCFace; each subject includes an HR image and 15 LR images. The results of visualization are shown after fine-tuning. As shown, stage one obtains preliminary clustering from chaotic distribution, and stage two significantly improves the compactness of clustering.

# Conclusion

In this paper, we propose a novel two-stage multi-scale resolution-adaptive (TMR) method for low-resolution face recognition. Unlike previous works, the proposed TMR method fully captures the HR-to-LR guidance and the correlations between the multiple LR images by introducing multi-scale distillation on intermediate features in the first stage and multi-resolution representation clustering in the second stage. Furthermore, a curriculum learning strategy is introduced for smooth training of the LR network. Experimental results on eight widely used benchmark databases demonstrate the superiority of our method. The effectiveness of each stage and the proposed module are also verified by comprehensive ablation studies and qualitative visualization.

# Reference

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. “A Simple Framework for Contrastive Learning of Visual Representations.” In *International Conference on Machine Learning*, 1597–1607. PMLR.

Cheng, Zhiyi, Xiatian Zhu, and Shaogang Gong. 2018a. “Low-Resolution Face Recognition.” In *Asian Conference on Computer Vision*, 605–21. Springer.

———. 2018b. “Surveillance Face Recognition Challenge.” *arXiv Preprint arXiv:1804.09691*.

Deng, Jiankang, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. “Arcface: Additive Angular Margin Loss for Deep Face Recognition.” In *Proceedings of the Ieee/Cvf Conference on Computer Vision and Pattern Recognition*, 4690–9.

Fang, Han, Weihong Deng, Yaoyao Zhong, and Jiani Hu. 2020. “Generate to Adapt: Resolution Adaption Network for Surveillance Face Recognition.” In *European Conference on Computer Vision*, 741–58. Springer.

Grgic, Mislav, Kresimir Delac, and Sonja Grgic. 2011. “SCface–surveillance Cameras Face Database.” *Multimedia Tools and Applications* 51 (3). Springer: 863–79.

He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. “Momentum Contrast for Unsupervised Visual Representation Learning.” In *Proceedings of the Ieee/Cvf Conference on Computer Vision and Pattern Recognition*, 9729–38.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. “Deep Residual Learning for Image Recognition.” In *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition*, 770–78.

Hjelm, R Devon, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. “Learning Deep Representations by Mutual Information Estimation and Maximization.” *arXiv Preprint arXiv:1808.06670*.

Huang, Gary B, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. “Labeled Faces in the Wild: A Database Forstudying Face Recognition in Unconstrained Environments.” In *Workshop on Faces in’Real-Life’Images: Detection, Alignment, and Recognition*.

Huang, Yuge, Pengcheng Shen, Ying Tai, Shaoxin Li, Xiaoming Liu, Jilin Li, Feiyue Huang, and Rongrong Ji. 2020. “Improving Face Recognition from Hard Samples via Distribution Distillation Loss.” In *European Conference on Computer Vision*, 138–54. Springer.

Jiao, Qianfen, Rui Li, Wenming Cao, Jian Zhong, Si Wu, and Hau-San Wong. 2021. “DDAT: Dual Domain Adaptive Translation for Low-Resolution Face Verification in the Wild.” *Pattern Recognition* 120. Elsevier: 108107.

Khalid, Syed Safwan, Muhammad Awais, Zhen-Hua Feng, Chi-Ho Chan, Ammarah Farooq, Ali Akbari, and Josef Kittler. 2020. “Resolution Invariant Face Recognition Using a Distillation Approach.” *IEEE Transactions on Biometrics, Behavior, and Identity Science* 2 (4). IEEE: 410–20.

Liu, Weiyang, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. “Sphereface: Deep Hypersphere Embedding for Face Recognition.” In *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition*, 212–20.

Low, Cheng-Yaw, Andrew Beng-Jin Teoh, and Jaewoo Park. 2021. “MIND-Net: A Deep Mutual Information Distillation Network for Realistic Low-Resolution Face Recognition.” *IEEE Signal Processing Letters* 28. IEEE: 354–58.

Lu, Ze, Xudong Jiang, and Alex Kot. 2018. “Deep Coupled Resnet for Low-Resolution Face Recognition.” *IEEE Signal Processing Letters* 25 (4). IEEE: 526–30.

Massoli, Fabio Valerio, Giuseppe Amato, and Fabrizio Falchi. 2020. “Cross-Resolution Learning for Face Recognition.” *Image and Vision Computing* 99. Elsevier: 103927.

Maze, Brianna, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, et al. 2018. “Iarpa Janus Benchmark-c: Face Data set and Protocol.” In *2018 International Conference on Biometrics (Icb)*, 158–65. IEEE.

Moschoglou, Stylianos, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. 2017. “Agedb: The First Manually Collected, in-the-Wild Age Database.” In *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition Workshops*, 51–59.

Parkhi, Omkar M, Andrea Vedaldi, and Andrew Zisserman. 2015. “Deep Face Recognition.” British Machine Vision Association.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. 2015. “Facenet: A Unified Embedding for Face Recognition and Clustering.” In *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition*, 815–23.

Sengupta, Soumyadip, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. 2016. “Frontal to Profile Face Verification in the Wild.” In *2016 Ieee Winter Conference on Applications of Computer Vision (Wacv)*, 1–9. IEEE.

Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. “Non-Local Neural Networks.” In *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Wen, Yandong, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. “A Discriminative Feature Learning Approach for Deep Face Recognition.” In *European Conference on Computer Vision*, 499–515. Springer.

Whitelam, Cameron, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, et al. 2017. “Iarpa Janus Benchmark-B Face Data set.” In *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition Workshops*, 90–98.

Yi, S., X. Wang, and X. Tang. 2014. “Deep Learning Face Representation by Joint Identification-Verification.” *Advances in Neural Information Processing Systems* 27.

Yin, Xi, Ying Tai, Yuge Huang, and Xiaoming Liu. 2020. “Fan: Feature Adaptation Network for Surveillance Face Recognition and Normalization.” In *Proceedings of the Asian Conference on Computer Vision*.

Zha, Juan, and Hongyang Chao. 2019. “Tcn: Transferable Coupled Network for Cross-Resolution Face Recognition.” In *ICASSP 2019-2019 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp)*, 3302–6. IEEE.