Dive into the Resolution Augmentations and Metrics in Low Resolution Face Recognition: A Plain yet Effective New Baseline

Xu Ling,[1] **Yichen Lu,**[1] **Wenqi Xu,**[1] **Weihong Deng,**[1] **Yingjie Zhang,**[1,2] **Xingchen Cui,**[1,2] **Hongzhi Shi,**[1,2] **Dongchao Wen** [*,1,2]

[1] Inspur Electronic Information Industry Co., Ltd
[2] Shandong Massive Information Technology Research Institute, Jinan, China
{lingxu421, wenqixu99}@gmail.com, yichen.iu@outlook.com, dengweihong2020@foxmail.com, {zhangyj-s, cuixingchen, shihzh, wendongchao}@inspur.com

Abstract

Although deep learning has significantly improved Face Recognition (FR), dramatic performance deterioration may occur when processing Low Resolution (LR) faces. To alleviate this, approaches based on unified feature space are proposed with the sacrifice under High Resolution (HR) circumstances. To deal with the huge domain gap between HR and LR domains and achieve the best on both domains, we first took a closer look at the impacts of several resolution augmentations and then analyzed the difficulty of LR samples from the perspective of the model gradient produced by different resolution samples. Besides, we also find that the introduction of some resolutions could help the learning of lower resolutions. Based on these, we divide the LR samples into three difficulties according to the resolution, and propose a more effective Multi-Resolution Augmentation. Then, due to the rapidly increasing domain gap as the resolution decreases, we carefully design a novel and effective metric loss based on a $LogExp$ distance function that provides decent gradients to prevent oscillation near the convergence point or tolerance to small distance errors; it could also dynamically adjust the penalty for errors in different dimensions, allowing for more optimization of dimensions with large errors. Combining these two insights, our model could learn more general knowledge in a wide resolution range of images and balanced results can be achieved by our extremely simple framework. Moreover, the augmentations and metrics are the cornerstones of LRFR, so our method could be considered a new baseline for the LRFR task. Experiments on the LRFR datasets: SCface, XQLFW and large-scale LRFR dataset: TinyFace demonstrate the effectiveness of our methods, while the degradation on HRFR datasets is significantly reduced.

1 Introduction

The high accuracy of High Resolution Face Recognition (HRFR) tasks mainly attributed to large-scale face data (Yi et al. 2014; Cao et al. 2018) collected in uncontrolled scenarios and various softmax-based loss functions (Liu et al. 2016, 2017; Wang et al. 2018; Deng et al. 2019). However, due to the huge domain gap between LR and HR faces and very limited LR faces data in the wild, the performance of the Low Resolution Face Recognition (LRFR) tasks is still poor. Furthermore, LRFR tasks are
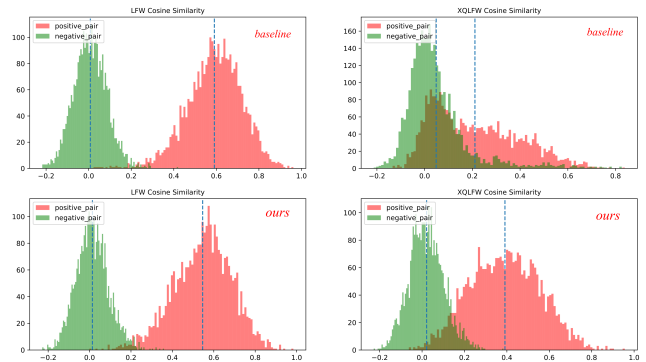
---
[*]Corresponding author

Figure 1: The negative and positive cosine similarity distribution of LFW (Huang et al. 2007) and XQLFW (Knoche, Hörmann, and Rigoll 2021) datasets, which represent the HR and LR cases correspondingly. In comparison to ours, we used the HR model as a baseline. For the LR case, our model can extract identity information from the faces, thus reducing the overlap of cosine similarity of positive and negative pairs. At the same time, the HR case has only been slightly harmed.

again widely seen in real scenarios, e.g. LR faces captured from surveillance cameras at long distances, which also push the problem to be improved (Fang et al. 2020).

As shown in Figure 1, if the HR model is directly applied to recognize LR faces, a large portion of the positive pairs will have close to orthogonal cosine similarity, while a small portion of the negative pairs will have very comparable features. This indicates that the information loss caused by the resolution and massive domain gap between LR and HR mislead the model to fail to extract the correct identity from the faces in the LR domain. Therefore, it becomes crucial to fuse the two widely disparate domains so that the model can be resolution-independent, and inadequate and inappropriate fusion can lead to undesirable side effects of reduced HR performance (Hong and Ryu 2019; Shi et al. 2020).

In general, data augmentation is the most common tool to resolve data insufficiency, and Multi-Resolution augmentations are commonly used in LRFR. However, previous work tend to simply select some resolutions for augmentation in equal proportions (ST-M (Knoche, Hörmann,

and Rigoll 2021):$7px$,$14px$,$28px$,$56px$, Bridge KD (Ge et al. 2020b):$16px$,$32px$,$64px$,$96px$, etc). They lack detailed study and analysis, and thus the settings may not be optimal. Therefore, we conduct some experiments to verify the importance of different resolutions. We first tested models with different resolution augmentation on synthetic LFW at different resolutions to analyze the effect of different resolution augmentation. More specifically, we found that the HR model could be robust to relatively high resolution inputs (from $28\ px$ to $112\ px$) and that even a $28\ px$ augmentation did not result in a significant performance gain. We also found that low-resolution augmentations can bring performance gains (from $7\ px$ to $14\ px$), however, due to the large gaps between domains, introducing them directly is often not optimal. At the same time, augmentation using multiple different resolutions gives better results than using only one resolution. Further, to maximum the potential of Multi-Resolution Augmentation (MAug), we use model back propagation gradient to analyze the learning difficulty of different resolution samples and then we give the relationship between the difficulty of the synthesized samples and the resolutions. In light of these, we classify the synthetic low-resolution faces into *extremely hard*, *hard* and *semi-hard* samples based on their resolution and propose a new multi-resolution augmentation. Significant performance improvements could be obtained.

Furthermore, from the model gradient we also found that there is learning migration between resolutions, i.e. the introduction of some resolutions in training could reduce the learning difficulty of samples with lower resolutions. Therefore, with proper Multi-Resolution Augmentation, a decent metric loss could leverage the gradient information to guide the model to gradually learn lower resolutions samples during training. We first analyzed the L_1 and L_2 losses from the gradient perspective. It may be found that due to the constant gradient value, L_1 loss leads to the oscillation problem as the two embeddings are unable to converge further when they are relatively close. On the other hand, L_2 loss fails to provide enough gradient when the distance is relatively close, which could lead to tolerance of tiny distance and insufficient convergence. These convergence difficulties can lead to inadequacies in fusing the two domains. So, we proposed a new metric loss L_{LogExp}, which could avoid neglecting the tiny distance and mitigate the oscillation problem, resulting in a more effective feature extractor.

In summary, the main contributions of this paper are as follows:

- We dive into the resolution augmentations to verify the importance of different resolutions and find the learning migration between resolutions.

- We introduce the concept of hard samples into resolution augmentation and propose a new multi-resolution augmentation method that only augments at low resolutions.

- We review the popular distance function from the perspective of gradient. Then, we propose a novel distance loss to better fuse the widely separated HR and LR domains, which is simple to implement and can be integrated with a variety of other approaches.

- Our proposed method can be considered a new baseline for the LRFR task, and it can be conveniently integrated into other methods. Meanwhile, we also conduct a large number of experiments and the state-of-the-art results in LRFR are achieved while the HRFR performance degradations are reduced.

2 Related Work

2.1 Common Face Recognition

Softmax-based loss (Deng et al. 2019; Wang et al. 2018; Liu et al. 2017, 2016), as well as metric learning loss (Schroff, Kalenichenko, and Philbin 2015; Sun 2015), have substantially improved general face recognition(or HR face recognition) (Wang and Deng 2018). Meanwhile, large-scale face datasets such as MillionCelebs (Zhang et al. 2020) and Webface42M (Zhu et al. 2021) serve as the foundations for face recognition to improve. However, since the detailed information in LR faces is less than in HR faces, resulting in a massive domain gap between HR and LR, the HR model's performance under LR circumstances would suffer significantly.

2.2 Low Resolution Face Recognition

SR-based Method. The Super Resolution(SR)-based approach expects to fill the lost information to generate synthetic HR images (Zangeneh, Rahmati, and Mohsenzadeh 2020; Chen et al. 2018; Ataer-Cansizoglu et al. 2019; Mudunuri, Sanyal, and Biswas 2018). To restrict the features of the synthetic images and HR images, Zhang et al. proposed Super-Identity loss. Fang et al. (Fang et al. 2020) suggested using MR-GAN to generate realistic LR images for training, avoiding the issue of unrealistic downsampling to some extent.

Feature-based Method. However, the quality of the synthesized HR images is difficult to be proven valid and may even carry the bias of the SR model, while the SR method increases the computational effort. Thus, on the other hand, feature-based works try to learn projection into a unified domain and minimize the distances between LR and HR embeddings. Lu, Jiang, and Kot proposed a deep coupled ResNet (DCR) model that employs branch networks to transform the image embeddings for each domain to learn robust features. Ge and Massoli et al. (Ge et al. 2020a,b, 2018; Zhao et al. 2019; Massoli, Amato, and Falchi 2020) expect to achieve better LR performance by using knowledge distillation to migrate helpful knowledge for LR from the HR model to the student network. Li, Tu, and Xu proposed using Rival Penalized Competitive Learning to drive the most similar class center away to learn more discriminative features, and high performance was attained with such a simple concept. Singh et al. recently proposed replacing the softmax loss with Derived-Margin Softmax Loss, which may alter the margin dynamically during training.

Previous Augmentations. Although LR faces can be obtained by GAN-based methods (Li, Tu, and Xu 2021) or Gaussian blurring methods (Shi et al. 2020), interpolation-based methods are the most popular (bicubic is most commonly-used method). They first down-
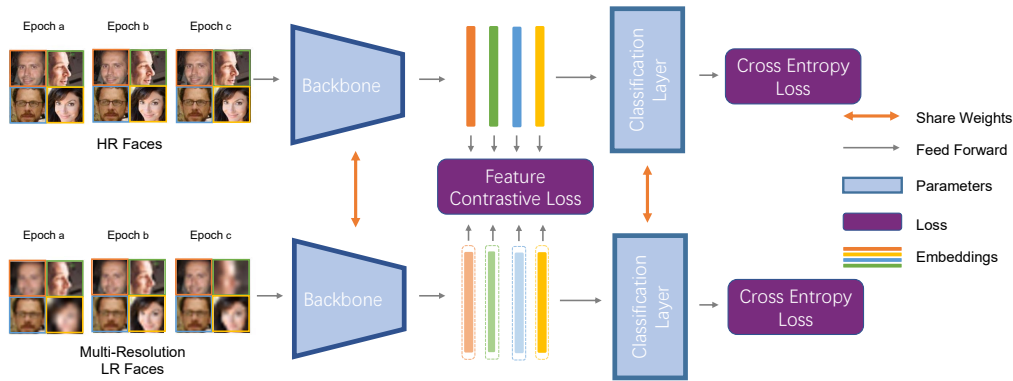
Figure 2: Overview of our proposed method. The current input HR faces and the multi-resolution LR faces are input into the backbone to get the corresponding embeddings. The Feature Contrastive Loss is used to constrain the distance. CosFace is used to ensure that embeddings are classified to the correct identity.

sampling the input to a certain resolution and then up-sampling it to fit the input dimension. Previous work uses a particular resolution (Zhao et al.:$32px$), or multi-resolution such as DCR:$40px$,$6px$, ST-M (Knoche, Hörmann, and Rigoll 2021):$7px$,$14px$,$28px$,$56px$, Bridge KD (Ge et al. 2020b):$16px$,$32px$,$64px$,$96px$, etc. However, Multi-Resolution Augmentation has not been carefully studied in the previous works. In this paper, we study in detail the relationship among resolution, performance and model gradient, and propose a novel and effective multi-resolution augmentation based on bicubic interpolation.

2.3 Contrastive Learning and Siamese Network

Our research takes a deep dive into the resolution augmentations and metric functions in LRFR and demonstrates their enormous potential. Therefore, we did not introduce a complex network structure, but resorted to the most naive structure in the LRFR method based on the unified feature space (Mudunuri, Sanyal, and Biswas 2018): the Siamese network structure. At the same time, the structure adopted in this paper is similar to contrastive learning. In self-supervised learning, contrastive-based algorithms have also achieved competitive results or even surpassed the performance of supervised learning in several tasks (Oord, Li, and Vinyals 2018; He et al. 2020; Chen et al. 2020). However, we does not introduce negative samples, which is the main difference from contrastive learning. In summary, we adopt the most intuitive, plain structure to demonstrate the effectiveness and superiority of our proposed technique.

3 Methodology

Guided by the Siamese network, our subsequent research and exploration are based on this most concise structure. The overview of our proposed method is illustrated in Figure 2. In this paper, we propose **1. a new Multi-Resolution Augmentation** and **2. a new Feature Contrast Loss**.

3.1 Multi-Resolution Augmentation

The resolution augmentations use interpolation method (e.g.BICUBIC) to downsample and then up-

sample HR images to reduce the image resolution but keep the input dimension. In this section, we mainly focus on three experiments to answer two questions: 1. In multi-resolution augmentation, which resolutions should be augmented, and 2. Why they are effective.

First, we analyze the effect of resolution augmentations when a feature contrast loss is used between the HR and the synthesized LR faces. As shown in Figure 3a, we test models with various resolution augmentations on synthetic LFW at different resolutions and give the SSIM Index (Wang et al. 2004) of different resolution images. Secondly, we compute the gradient sums for different models in Figure 3a for different inputs, as shown in Figure 3b. We also provide the t-SNE (Van der Maaten and Hinton 2008) results of the HR model for the features of the same faces at 7 representative different resolutions, as shown in Figure 3c. Accordingly, we classify the difficulty of different resolution samples and propose a new Multi-Resolution Augmentation.

As the 112 px could be the most common input size in FR. 7 px and 14 px, as the 16x and 8x scaling of 112 px, is widely considered in Multi-Resolution Augmentations (Knoche, Hörmann, and Rigoll 2021). However, as for the 28 px, which was also selected as one of the augmentation resolutions by previous work, we argue that its improvement in performance is marginal. It could be found from Figure 3a that the HR model is robust to relatively high resolution images (from 28 px to 112 px) and its performance only starts to drop dramatically when the resolution is lower than 28 px. In addition, the use of 28 px resolution augmentation, as shown in Figure 3a, does not greatly help the model to improve its performance. The same conclusion can be drawn from Figure 3c, where the features are significantly shifted starting from 20 px and up to 14 px and 7 px. These indicate that relatively high resolution images are not completely unfamiliar to the HR model, which we believe mainly thank to the presence of some LR images in the original large-scale training data.Therefore, the relatively high resolution augmentations (bigger than 28 px) commonly used in previous methods do not need to be introduced.

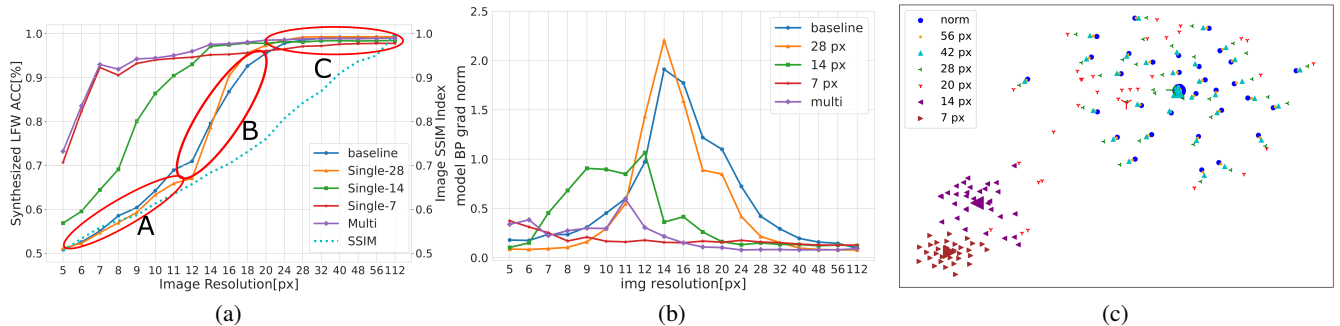However, for the introduction of augmentations with a

Figure 3: (a):Synthesized LFW accuracy for different image resolutions with our L_{LogExp}. *Single-x* denotes that only x px and HR faces are utilized to train the model. *Multi* indicates that multiple resolution images are used during training where we employed the 7 px,14 px,20 px and HR faces, while the training set size remains the same as *Single-x*. The general HR model is used as *baseline*. The SSIM Index for different resolutions are also shown. (b):The gradient norm sum of the parameters of the model in (a) after back-propagation for different resolution inputs. (c):The t-SNE result of the HR model (Baseline in (a)) for the features of the same faces at different resolutions, the bigger point represents the center of each resolution.

resolution of 7 px or 14 px alone, although performance gains at low resolutions are obtained as shown in Figure 3a, performance deterioration occurs at high resolutions compared to the HR model. Similarly, the feature points in Figure 3c with resolution above 28 px could always find their corresponding HR faces, while when the resolution is less than 20 px, this correspondence of the same image is not obvious. And even the 14 px and 7 px features form their own clusters. These indicates that the loss of information increases rapidly as the resolution decreases; meanwhile, the domain gap between low resolutions may be greater than those between low and high resolutions, so the direct introduction of low resolution will make the HR performance undergo a serious degradation.

One option for improving the performance of resolution augmentation is to use multiple resolutions simultaneously, and we here give a possible explanation of why multi-resolution augmentation is more effective. When inputting different samples, the model will derive different gradients to optimize the loss, and the norm of the gradient can determine the magnitude of the amount of change the samples make to the model parameters. The norm of the gradient is also often used in hard mining (Li, Liu, and Wang 2019). As shown in Figure 3b, the gradient of the models shows a trend of increasing and then decreasing with decreasing resolution. Also, we found some migration between resolutions: the introduction of some resolutions shifted the peak of the gradients to a smaller resolution. This indicates that the introduction of some resolutions will make it less difficult to learn at lower resolutions (7px-14px is much easier for the Single-14 model than Baseline, as in Figure 3a&3b). That's why Multi-Resolution could obtain a better result. Different but proper resolution could be the bridge to fuse the HR and LR domains well.

On the other hand, due to the huge domain gap between 14 px and 112 px, directly introducing 7 px and 14 px as Multi-Resolution Aug would cause the model to suffer too much on HR (as shown in Table 6); at the same time, due to

the aforementioned migration between resolutions, choosing a resolution between 14 px and 112 px for augmentation could better help the model's learning. It is obvious from Figure 3b that 20 px is the point where the feature center starts to shift significantly. So based on the above three points, we also introduce 20 px into our multi-resolution augmentation as well.

Finally, as for the learning difficulty of different resolutions, one could figure out from Figure 3a&3b that that the model gradient and performance are negatively correlated in the 14px-112px region. However, when the resolution is less than 14 px, the gradient of the model is small and the performance is also low in Figure 3a, indicating that this part of the samples are extremely hard samples which cannot produce enough gradients, making optimization challenging. Therefore based on our experiments and findings above, we argue that samples below 12 px should be considered as *extremely hard* samples (A in Figure 3a), and combined with Figure 3c, one can consider 20 px to 32 px as *semi-hard* samples (C in Figure 3a), and 12 px to 20 px as *hard* samples (B in Figure 3a). Thus, we propose to combine three different difficulties in training, using three representative resolutions of 7 px, 14 px, and 20 px for augmentation.

3.2 Feature Contrast Loss

Due to the huge domain gap between LR and HR domains, the feature contrast loss need to be carefully designed. However, we find that the commonly-used L_1 and L_2 losses exhibit similar performance under our framework (As shown in Table 4), which indicates that neither is optimal. We thus propose a novel Feature Constraint Loss L_{LogExp} which incorporates the advantages of both based on the analysis of their merits.

Analysis of Existed Distance Functions. First, for x and y are the two vectors we want to pull together, L_1 loss:

$$L_1(x, y) = \frac{1}{D} \sum_{i=1}^{D} |x_i - y_i|, \qquad \frac{\partial L_1}{\partial x_i} = \frac{1}{D} \cdot (\pm 1), \quad (1)$$
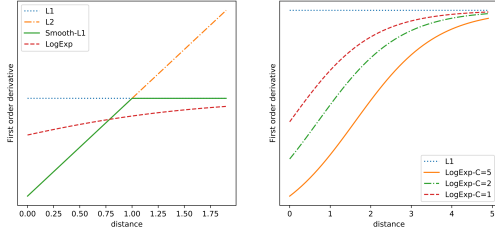
4

Figure 4: Left:The First-order derivative values of the four distance functions. Right:The effect of different C on the first-order derivatives of the L_{LogExp} functions in Equation 6. Our L_{LogExp}, could not only provide different gradients according to the error on the current dimension (Left), but also adjust the gradient magnitude according to the ratio of errors between dimensions (Right).

gives the same penalty to all distance errors in all dimensions ($|\frac{\partial L_1}{\partial x_i}| = \frac{1}{D}$). The errors in different dimensions tend to be different. Therefore, when we have optimized the dimensions with relatively small errors by L_1, the dimensions with large errors need further optimization, which leads to the oscillation problem near the convergence point. On the other hand, although L_2 loss:

$$L_2(x,y) = \frac{1}{2D}\sum_{i=1}^{D}(x_i - y_i)^2, \qquad \frac{\partial L_2}{\partial x_i} = \frac{1}{D}\cdot(x_i - y_i),$$
$$(2)$$

treats different dimensions based on their own distance errors, tolerance of small distance errors may occur because of the small gradient of L_2 ($|\frac{\partial L_2}{\partial x_i}| = \frac{1}{D}\cdot|x_i - y_i|$) when the two embeddings are relatively close, which is also disappointed. The $SmoothL_1$ function, as a combination of L_1 and L_2 would also encounter the same problem as L_2.

Our $LogExp$ **Distance Function.** Therefore, based on the analysis above, a distance function that not only provides different gradients considering the errors in different dimensions (Advantage of L_2), but also provides gradients with sufficient magnitude when the two embeddings are close to one another (Advantage of L_1), would be the proper one. Additionally, due to the direct constraint between embeddings, we also expect its gradient to be smooth. Here we give a novel Feature Constraint Loss L_{LogExp} with the formula and gradient:

$$L_{LogExp} = \frac{1}{D}\log(1 + \sum_{i=1}^{D}(e^{|x_i - y_i|} - 1)), \qquad (3)$$

$$\frac{\partial L_{LogExp}}{\partial x_i} = \frac{1}{D}\cdot\frac{(\pm 1)\cdot e^{|x_i - y_i|}}{1 + \sum_{i=1}^{D}(e^{|x_i - y_i|} - 1)}. \qquad (4)$$

The first-order gradient of L_{LogExp} is shown in Figure 4. Also, just like both L_1 and L_2 could be generalized as: $L_p = \frac{1}{p\cdot D}\sum_{i=1}^{D}(|x_i - y_i|)^p$, L_p Loss is L_1 loss when $p = 1$ and is L_2 loss when $p = 2$, we could also reference this idea to

give L_{LogExp} in general form as:

$$L_{LogExp} = \frac{1}{p\cdot D}\log(1 + \sum_{i=1}^{D}(e^{(|x_i - y_i|)^p} - 1)). \qquad (5)$$

More Advantage of L_{LogExp}**.** The above gradient formula Equation 4, could be rewritten by considering the terms unrelated to the current dimension as a remainder term C, and thus can be rewritten as :

$$\frac{\partial L_{LogExp}}{\partial x_i} = \frac{1}{D}\cdot\frac{(\pm 1)\cdot e^{|x_i - y_i|}}{C + e^{|x_i - y_i|}}. \qquad (6)$$

As shown in Equation 4 and Figure 4, our novel distance function L_{LogExp} gives an intermidiate gradient between the L_1 and L_2.Similarly, the remainder term C in Equation 6 represents the sum of distance in the other dimensions, and the effect of C is shown in Figure 4. When C is large, the gradient value of the current dimension becomes smaller, while when C is small, the gradient value of the current dimension becomes larger. This can be considered that the L_{LogExp} function could dynamically adjust the gradient magnitude according to the ratio of errors between dimensions.Thus, it is possible to focus on the dimensions with larger errors during the training process to improve this part first.

3.3 Total Loss

Cross-entropy loss with class label can push negative pairs apart, thus we only employ contrast loss to pull positive pairs together.Therefore, we could simply use the contrast loss to constrain the distance between the corresponding HR and LR faces and make them classify to the right class with the help of Softmax-based loss, e.g. CosFace (Wang et al. 2018), to learn discriminative knowledge.

So, our proposed training process and formulation are as follows: I_{HR} represents the current input HR face from a large-scale dataset, e.g. WebFace (Yi et al. 2014). First, a resolution is randomly selected from the multi-resolution list and then I_{HR} is augmented with BICUBIC to obtain the corresponding I_{LR}. Thereafter, a convolutional neural network (CNN) backbone (He et al. 2016) is used to obtain the embeddings f of I_{HR} and I_{LR}, $f_{HR} = F_{CNN}(I_{HR})$ and $f_{LR} = F_{CNN}(I_{LR})$. Then CosFace (Wang et al. 2018) is used as the classification loss L_{cls} and combines the proposed contrastive loss L_{LogExp} to constrain the two embeddings after normalization. The final total loss obtained is

$$L_{dist} = L_{LogExp}(\frac{f_{HR}}{||f_{HR}||_2}, \frac{f_{LR}}{||f_{LR}||_2}),$$
$$L_{all} = \lambda\cdot L_{dist} + \frac{1}{2}L_{cls}(f_{HR}) + \frac{1}{2}L_{cls}(f_{LR}). \qquad (7)$$

4 Experiments

4.1 Dataset and Training Settings

Dataset. We chose the large-scale face datasets CASIA-WebFace (Yi et al. 2014) and the cleaned version (Deng et al. 2019) of MS1M (Guo et al. 2016) as training dataset. For

Methods	$d1$	$d2$	$d3$	Avg
DCR (2018)	73.3	93.5	98.0	88.27
TCN (2019)	74.6	94.9	98.6	89.37
DATL (2019)	76.2	96.9	98.1	90.40
DAQL (2019)	77.3	96.6	98.1	90.66
RAN (2020)	81.3	97.8	98.8	92.63
DDL(Res50) (2020)	86.8	98.3	98.3	94.47
NPT-Loss(Res50) (2022)	85.7	*99.1*	99.1	94.63
RPCL-DCR (2022)	90.4	98.0	98.0	95.47
Ours(w/o FT)	73.0	96.0	98.8	89.27
Ours(Res34)	*91.8*	99.0	*99.3*	*96.67*
Ours(Res50)	**92.5**	**99.3**	**99.5**	**97.08**

Table 1: Rank-1 face identification accuracy (%) on the SCface (Grgic, Delac, and Grgic 2011); 'w/o FT' denotes straight testing with the trained model without fine-tuning.

the test set, SCface (Grgic, Delac, and Grgic 2011) is used as the common cross-resolution recognition dataset, which contains 130 ids and 4,160 face images. XQLFW (Knoche, Hörmann, and Rigoll 2021) is a recently proposed challenging cross-resolution dataset that solves the problem of non-uniformity of the synthetic LFW datasets used in previous work (Li, Tu, and Xu 2021). TinyFace (Cheng, Zhu, and Gong 2018) is also used to test the performance on large-scale low resolution faces, which contains 169,403 LR faces. We also used LFW (Huang et al. 2007) to measure the performance degradation under HR scenarios.

Implementation Details. In the training process, the resolution and ratio of 7 px: 14 px: 20 px = 1:1:2 are used for multi-resolution augmentation. Our resolution augmentation is based on BICUBIC. Random Horizontal Flip is used for data augmentation. All training data were aligned and cropped to 112×112 utilize MTCNN (Zhang et al. 2016). SGD momentum was used as the optimizer for training and the momentum was set to 0.9, and weight decay was 5e-4. When CASIA-WebFace (Yi et al. 2014) was used as training set, ResNet34 (He et al. 2016) was used as the backbone network with initial learning rate was set to 0.1, a total of 28 epochs were trained and the learning rate was divided by 10 at 16 and 24 epochs. When MS1M was used as the training set, ResNet50 (He et al. 2016) was used as the backbone, the initial learning rate was set to 0.01, a total of 16 epochs are trained and the learning rate is divided by 10 at 8 and 12 epochs. The λ is set to 1. The s in CosFace loss is set to 48 and m is set to 0.4. We perform our experiments on two GTX1080TIs, the implementation is based on PyTorch (Paszke et al. 2019) and the batch size is set to 256.[1]

4.2 Compare With Other Methods

Results On SCface. Following (Lu, Jiang, and Kot 2018; Li et al. 2019; Yang et al. 2017) 80 of the 130 ids were used for testing, and the remaining 50 ids that did not intersect were utilized to fine-tune the entire network. The results of our method and the comparison with the previous method

---

[1]The code is released at: https://github.com/hurricanelx/LRFR

Methods	R-1	R-5	R-10	R-20
DCR (2018)	0.29	-	0.40	0.44
PeiLi's (2018)	0.31	-	0.43	0.46
Cheng's (2018)	0.36	-	0.46	0.50
CosFace(Res34)	0.29	-	0.39	0.42
RPCL(Res34) (2022)	0.32	-	0.41	0.44
CosFace(Res100)	0.47	0.52	-	-
URL(Res100) (2020)	**0.63**	**0.68**	-	-
Ours(Res34)	0.39	0.46	0.50	0.54
Ours(Res50)	*0.59*	*0.64*	**0.68**	**0.73**

Table 2: Face identification Rank-1, Rank-5, Rank-10 and Rank-20 accuracy on TinyFace. R-1 represent for Rank-1.

Models	LFW	XQLFW	Avg
CosFace(Res34) (2018)	99.42	68.95(-30.47)	84.19
ArcFace(Res50) (2019)	99.50	74.22(-25.28)	86.86
MagFace(Res50) (2021)	99.63	76.95(-22.68)	88.29
QMagFace-50 (2021)	**99.73**	80.63(-19.1)	90.18
BT-M(Res50) (2021)	99.30	83.6(-15.70)	91.45
ST-M2(Res50) (2021)	95.87	90.82(-5.05)	93.35
Face Transformer (2021)	99.70	87.9(-11.80)	93.80
ST-M1(Res50) (2021)	97.30	90.97(-6.33)	94.14
Ours(Res34)	99.03	92.18(-6.85)	95.61
Ours(Res50)	99.40	**94.33**(-5.07)	**96.87**

Table 3: Face verification accuracy (%) for several previous approaches on LFW and XQLFW. Most compared results are directly cited from (Knoche, Hörmann, and Rigoll 2021).

are shown in Table 1. From this, it could figure out that, after fine-tuning, our method is superior to all prior methods, and reaches the state-of-the-art results. It's also worth noting that, our Res34-based model outperforms DDL (Huang et al. 2020) and NPT-Loss (Khalid et al. 2022) despite having ResNet50 as their backbone. Furthermore, the performance of our method without fine-tuning is slightly outperformed the DCR (Lu, Jiang, and Kot 2018) technique, i.e., our model could do even better than DCR (Lu, Jiang, and Kot 2018) despite not having seen the SCface data. As for the generative based method (Fang et al. 2020), our method could also demonstrate better performance.

Results On TinyFace. TinyFace (Cheng, Zhu, and Gong 2018) is a large-scale low resolution face dataset of 169,403 LR faces that were captured under surveillance conditions. Following the same setting with previous work for a fair comparison, we use the training set of TinyFace to fine-tune our trained model before 1:N identification tests. The test results are shown in Table 2. Our method provides outstanding results, RPCL enhances CosFace's result by 3%, while our method improves it by 10%. It's also worth noting that despite having much more complex loss functions and twice as many parameters, the URL's (Shi et al. 2020) result is only 4% better than our ResNet50-based method.

Results On XQLFW. The results are presented in Table 3. From the results, for previous LRFR methods, e.g. ST-M1, ST-M2 (Knoche, Hörmann, and Rigoll 2021), although it
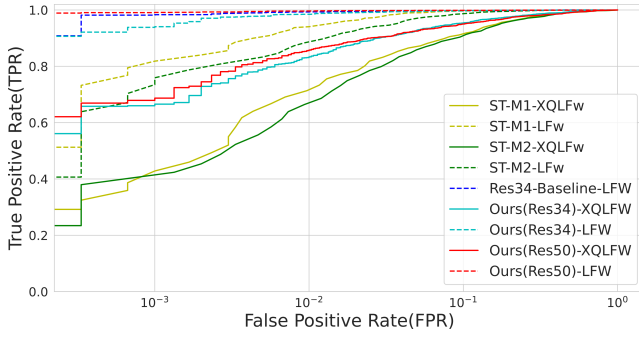
Figure 5: ROC curve of the previous model and our methods on LFW (Huang et al. 2007) and XQLFW (Knoche, Hörmann, and Rigoll 2021).

Methods	SCface w/o FT				XQLFW	LFW
	d1	d2	d3	Avg		
Baseline	40.3	92.8	98.5	77.2	68.95	**99.42**
Aug	71.5	92.8	96.3	86.8	84.33	98.65
MAug	67.0	95.3	98.0	86.8	90.89	98.98
PAug+Ours(1)	60.3	95.3	**98.8**	84.8	90.35	99.02
Aug+L_1	67.3	94.5	94.8	85.5	85.23	98.57
Aug+Ours(1)	70.0	94.8	96.5	87.1	87.78	98.60
MAug+Ours(1)	73.0	**96.0**	**98.8**	**89.3**	**92.18**	99.03
Aug+L_2	71.8	90.0	92.8	84.8	84.87	98.50
Aug+Ours(2)	72.5	92.3	94.3	86.3	86.88	98.63
MAug+Ours(2)	**74.0**	95.5	97.8	*89.1*	*91.98*	*99.08*

Table 4: Ablation Study based on ResNet34 and CASIA-WebFace. *Baseline* is the HR model with CosFace. *Aug* indicates that we only introduced $14\ px$ BICUBIC-based LR faces during training. *MAug* indicates that Multi-resolution augmentation is used in training. *PAug* indicates that the augmentation used in ST-M ($7px$,$14px$,$28px$,$56px$) is used for comparison. L_1 and L_2 represent that we introduced L_1 and L_2 losses to constrain the embedding of a pair of LR and HR faces, respectively. Ours(x) indicates that we use L_{LogExp} with $p = x$ as a replacement for L_1 or L_2.
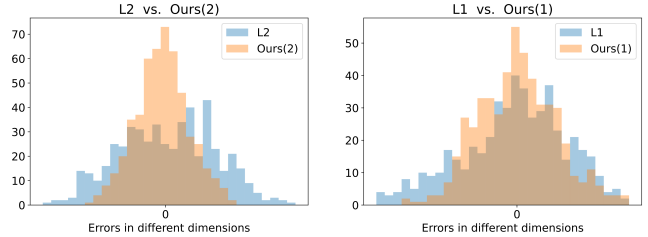

Figure 6: Distribution of the mean values of the errors in different dimensions of the HR and LR features. All four models are based on our *MAug* method.

improves the performance on the XQLFW dataset significantly compared to HR methods (ArcFace) (from 75% to 91%), it also faces the degradation on LFW (from 99.5% to 96%-97.3%). However, as demonstrated in Table 3, our Res50-based method improves accuracy by 20.11% on XQLFW compared to ArcFace baseline and just decays 0.10% on LFW. In contrast, compared to the prior technique ST-M1 (Knoche, Hörmann, and Rigoll 2021), which reached the optimum on XQLFW, our method improves 2.31% on XQLFW and 3.36% on LFW simultaneously. That is, our approach could improve the performance of the LRFR task while leaving the LFW performance mostly unaffected, which is both difficult and essential.

Figure 5 illustrates the ROC curve of the previous model and our methods on LFW and XQLFW. Similar conclusions could be drawn from Figure 5 as from Table 3. Meanwhile, our method could achieve a much higher True Positive Rate at low False Positive Rates compared to the previous method, e.g.ST-M (Knoche, Hörmann, and Rigoll 2021), which is a very significant improvement and also very meaningful. It also might be seen that the results of the ResNet50-based method could obtain a higher performance than the ResNet34-based method (0.37% on LFW and 2.15% on XQLFW). Our approach demonstrates consistency for the enhancement of backbone and dataset size, which is very meaningful and practical.

4.3 Ablation Study

The findings of our ablation study on SCface, XQLFW, and LFW may be found in Table 4. The performance improvement of our proposed *MAug* and L_{LogExp} is consistent and the improvement of each module is very significant

The effectiveness of L_{LogExp}. It could be observed that the introduction of a distance function (L_1, L_2) to constrain the distance between a pair of face embeddings would help the model to improve its performance on LRFR (about 2% on XQLFW). However, some further performance deterioration on LFW is also observed. Moreover, after replacing L_1 and L_2 with L_{LogExp}, our approach could improve performance on LRFR datasets while reducing performance decay on HRFR. For the case of $p = 1$, the increase is 2.55% on

XQLFW, while for the case of $p = 2$, it is 2% (Line 5,6; 8,9). With the introduction of Multi-Resolution augmentation, L_{LogExp} can also bring performance gains (Line 3,7,10) (3% on SCface and 1% on XQLFW, while almost unaffected on LFW). This indicates that introducing direct constraints on features could help the model learn uniform knowledge between LR and HR, but inappropriate constraints may lead to performance degradation on HR, however, our novel distance function could overcome this dilemma.

We also selected a set of HR faces, generated the corresponding LR faces, and calculated the distribution of the mean values of the different dimensional errors after extracting features based on four different models. All four models are based on our *MAug* method while using L_1, L_2, L_{LogExp} ($p = 1$) and L_{LogExp} ($p = 2$), respectively. As shown in Figure 6, the error of our L_{LogExp} is much smaller and can be considered significantly better than the results of L_1 and L_2 losses.

The effectiveness of *MAug*. Our new Multi-Resolution augmentation strategy could also bring performance improvements. With and without introducing constraint functions, ours *MAug* show performance improvements over *Aug*
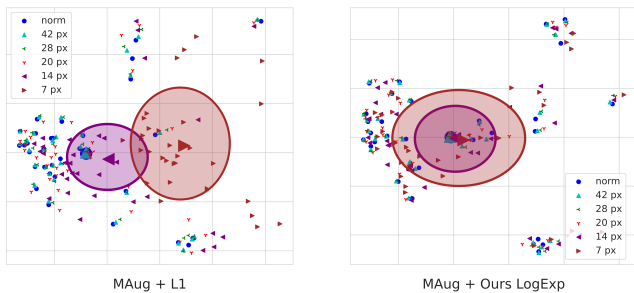
Figure 7: Feature PCA results for models with different contrast loss for the same set of faces at different resolutions (HR,$42px$,$28px$,$20px$,$14px$,$7px$,). Left: L_1, Right: Ours L_{LogExp}. These two models both utilized our *MAug* method. The bigger point represents the center.

λ	0.1	1	3	5	7	10
XQLFW	91.97	92.18	92.25	**92.45**	92.31	92.17
LFW	98.86	**99.03**	98.91	98.89	98.91	98.78

Table 5: The Impact of λ on LFW and XQLFW.

in both LRFR as well as HRFR (Line 2,3; 6,7; 9,10). Significant performance improvements can also be found when comparing with the augmentation (*PAug*) in ST-M (Knoche, Hörmann, and Rigoll 2021) (Line 4,7) (1.5% on XQLFW and 4.4% on SCface). These indicates that Multi-Resolution augmentation could help the model learn more general knowledge in both LR and HR, helping it to handle the differences between resolution domains and then project a wide range of resolution images to a unified feature space. This is consistent with our findings in Section 3.1 and therefore supports our conclusion.

Visualization Analysis. To show the effectiveness and superiority of our L_{LogExp} more intuitively, we performed a visualization analysis, and the results are shown in Figure 7. We selected a set of HR faces and generated corresponding LR faces with 5 different resolutions, after which we observed the distribution of features by PCA (Pearson 1901). Our proposed L_{LogExp} can indeed optimize the model better and fuse the HR and LR domains better, resulting in a better representation.

The impact of λ. We also performed more detailed ablation study on λ. As shown in Table 5, the impact of λ on performance is not significant. Larger λ means stricter constraints, which will bring performance improvement on XQLFW but performance degradation on LFW. However, excessively strict constraints will harm both. The optimal choice on LR is $\lambda = 5$, however in the main text we choose $\lambda = 1$ to minimize HR performance degradation, which could also achieve considerable LR performance.

Analysis on *MAug* Ratio. We also investigate the ratio of different resolutions in multi resolution augmentation as shown in Table 6. The results in Lines 3-5 illustrate that the different resolutions we introduced ($7px$;$14px$;$20px$) are all

Ratio	SCface w/o FT				XQLFW	LFW
	$d1$	$d2$	$d3$	Avg		
Baseline	40.3	92.8	98.5	77.2	68.95	**99.42**
MAug only	67.0	95.3	98.0	86.8	90.89	98.98
1:1:0	**73.8**	94.0	96.8	88.2	**92.43**	98.68
0:1:1	71.5	95.8	98.3	88.5	83.52	99.05
1:0:1	65.3	94.8	97.8	86.0	90.53	99.02
1:1:1	72.5	96.5	98.3	89.1	91.45	98.88
1:2:1	72.8	95.0	98.0	88.6	91.68	98.82
2:1:1	72.8	**96.8**	97.3	88.9	92.05	98.75
1:1:2	73.0	96.0	**98.8**	89.3	92.18	99.03

Table 6: Ablation Study of the Multi-Resolution Augmentation ratio ($7px$:$14px$:$20px$) based on ResNet34 and L_{LogExp}.

effective and indispensable. This also proves that the difficulty division for resolutions proposed by us is reasonable. The lack of *extremely hard* ($7px$) or *hard* ($14px$) samples will result in unsatisfactory LR performance (SCface $d1$; XQLFW), while the lack of *semi-hard* samples ($20px$) will cause a fragmentation of the domain between LR and HR and affect the performance under HR (SCface $d3$, LFW). Only by combining resolutions of different difficulty, multi-resolution augmentation can exert the maximum performance.

Moreover, similar to the findings of FaceNet (Schroff, Kalenichenko, and Philbin 2015), we experimentally prove that among these three different resolutions, we need to pay more attention to the *semi-hard* samples (Lines 7-9). The introduction of *semi-hard* samples $20px$ is to better build a bridge between the LR and HR domains, so as to optimize the hard samples more seamlessly. However, if pay unnecessarily too much attention to the *hard* and *extremely hard* samples, it will lead to poor performance due to optimization difficulties. Therefore, we finally adopt the ratio of 1:1:2 and verify its performance.

5 Discussion and Conclusion

In this paper, we dive into the resolution augmentations and the metrics in LRFR. We first analyze the impact of different resolution augmentations. Combining the analysis of model gradients, we suggest dividing the low resolution into three difficulties and propose a new Multi-Resolution Augmentation. We also analyze the popular L_1 and L_2 distance functions from a gradient viewpoint and propose a novel distance function L_{LogExp} that combines the benefits of both. By combining these two techniques, our model could learn more general knowledge in both LR faces and HR faces and boost the performance simultaneously.

Potential Social Impacts. Our experiments use the MS1M dataset, which has been withdrawn by its creator. However, our usage of MS1M is necessary for a pair comparison to previous works. Moreover, the subject of this paper is Low Resolution Face Recognition, which may be used in monitoring facilities to obtain face data or for surveillance purposes, etc., without people's knowledge.

References

Ataer-Cansizoglu, E.; Jones, M.; Zhang, Z.; and Sullivan, A. 2019. Verification of very low-resolution faces using an identity-preserving deep face super-resolution network. *arXiv preprint arXiv:1903.10974*.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; and Yang, J. 2018. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2492–2501.

Cheng, Z.; Zhu, X.; and Gong, S. 2018. Low-resolution face recognition. In *Asian Conference on Computer Vision*, 605–621. Springer.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Fang, H.; Deng, W.; Zhong, Y.; and Hu, J. 2020. Generate to adapt: Resolution adaption network for surveillance face recognition. In *European Conference on Computer Vision*, 741–758. Springer.

Ge, S.; Zhang, K.; Liu, H.; Hua, Y.; Zhao, S.; Jin, X.; and Wen, H. 2020a. Look one and more: Distilling hybrid order relational knowledge for cross-resolution image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10845–10852.

Ge, S.; Zhao, S.; Li, C.; and Li, J. 2018. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing*, 28(4): 2051–2062.

Ge, S.; Zhao, S.; Li, C.; Zhang, Y.; and Li, J. 2020b. Efficient low-resolution face recognition via bridge distillation. *IEEE Transactions on Image Processing*, 29: 6898–6908.

Ghosh, S.; Singh, R.; and Vatsa, M. 2019. On learning density aware embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4884–4892.

Grgic, M.; Delac, K.; and Grgic, S. 2011. SCface–surveillance cameras face database. *Multimedia tools and applications*, 51(3): 863–879.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Msceleb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, 87–102. Springer.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hong, S.; and Ryu, J. 2019. Unsupervised face domain transfer for low-resolution face recognition. *IEEE Signal Processing Letters*, 27: 156–160.

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.

Huang, Y.; Shen, P.; Tai, Y.; Li, S.; Liu, X.; Li, J.; Huang, F.; and Ji, R. 2020. Improving face recognition from hard samples via distribution distillation loss. In *European Conference on Computer Vision*, 138–154. Springer.

Khalid, S. S.; Awais, M.; Feng, Z.; Chan, C. H.; Farooq, A.; Akbari, A.; and Kittler, J. 2022. NPT-Loss: Demystifying face recognition losses with Nearest Proxies Triplet. *IEEE transactions on pattern analysis and machine intelligence*.

Knoche, M.; Hörmann, S.; and Rigoll, G. 2021. Image Resolution Susceptibility of Face Recognition Models. *arXiv preprint arXiv:2107.03769*.

Knoche, M.; Hörmann, S.; and Rigoll, G. 2021. Cross-Quality LFW: A Database for Analyzing Cross-Resolution Image Face Recognition in Unconstrained Environments. arXiv:2108.10290.

Li, B.; Liu, Y.; and Wang, X. 2019. Gradient harmonized single-stage detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8577–8584.

Li, P.; Prieto, L.; Mery, D.; and Flynn, P. 2018. Face recognition in low quality images: a survey. *arXiv preprint arXiv:1805.11519*.

Li, P.; Prieto, L.; Mery, D.; and Flynn, P. J. 2019. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, 14(8): 2000–2012.

Li, P.; Tu, S.; and Xu, L. 2021. Deep Rival Penalized Competitive Learning for Low-resolution Face Recognition. *arXiv preprint arXiv:2108.01286*.

Li, P.; Tu, S.; and Xu, L. 2022. Deep Rival Penalized Competitive Learning for low-resolution face recognition. *Neural Networks*.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 212–220.

Liu, W.; Wen, Y.; Yu, Z.; and Yang, M. 2016. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, 7.

Lu, Z.; Jiang, X.; and Kot, A. 2018. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(4): 526–530.

Massoli, F. V.; Amato, G.; and Falchi, F. 2020. Cross-resolution learning for face recognition. *Image and Vision Computing*, 99: 103927.

Meng, Q.; Zhao, S.; Huang, Z.; and Zhou, F. 2021. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14225–14234.

Mudunuri, S. P.; Sanyal, S.; and Biswas, S. 2018. GenLR-Net: Deep framework for very low resolution face and object recognition with generalization to unseen categories. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 602–60209. IEEE.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11): 559–572.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Shi, Y.; Yu, X.; Sohn, K.; Chandraker, M.; and Jain, A. K. 2020. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6817–6826.

Singh, M.; Nagpal, S.; Singh, R.; and Vatsa, M. 2021. DeriveNet for (Very) Low Resolution Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sun, Y. 2015. *Deep learning face representation by joint identification-verification*. The Chinese University of Hong Kong (Hong Kong).

Terhörst, P.; Ihlefeld, M.; Huber, M.; Damer, N.; Kirchbuchner, F.; Raja, K.; and Kuijper, A. 2021. QMagFace: Simple and Accurate Quality-Aware Face Recognition. *arXiv preprint arXiv:2111.13475*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, F.; Cheng, J.; Liu, W.; and Liu, H. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7): 926–930.

Wang, M.; and Deng, W. 2018. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Yang, F.; Yang, W.; Gao, R.; and Liao, Q. 2017. Discriminative multidimensional scaling for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(3): 388–392.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Zangeneh, E.; Rahmati, M.; and Mohsenzadeh, Y. 2020. Low resolution face recognition using a two-branch deep convolutional neural network architecture. *Expert Systems with Applications*, 139: 112854.

Zha, J.; and Chao, H. 2019. Tcn: Transferable coupled network for cross-resolution face recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3302–3306. IEEE.

Zhang, K.; Zhang, Z.; Cheng, C.-W.; Hsu, W. H.; Qiao, Y.; Liu, W.; and Zhang, T. 2018. Super-identity convolutional neural network for face hallucination. In *Proceedings of the European conference on computer vision (ECCV)*, 183–198.

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

Zhang, Y.; Deng, W.; Wang, M.; Hu, J.; Li, X.; Zhao, D.; and Wen, D. 2020. Global-Local GCN: Large-Scale Label Noise Cleansing for Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, S.; Gao, X.; Li, S.; and Ge, S. 2019. Low-resolution face recognition in the wild with mixed-domain distillation. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 148–154. IEEE.

Zhong, Y.; and Deng, W. 2021. Face Transformer for Recognition. *arXiv preprint arXiv:2103.14803*.

Zhu, Z.; Huang, G.; Deng, J.; Ye, Y.; Huang, J.; Chen, X.; Zhu, J.; Yang, T.; Lu, J.; Du, D.; et al. 2021. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10492–10502.