Learning Unified Representations for Multi-Resolution Face Recognition

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work, we propose Branch-to-Trunk network (BTNet), a novel representation learning method for multi-resolution face recognition. It consists of a trunk network (TNet), namely a unified encoder, and multiple branch networks (BNets), namely resolution adapters. As per the input, a resolution-specific BNet is used and the output are implanted as feature maps in the feature pyramid of TNet, at a layer with the same resolution. The discriminability of tiny faces is significantly improved, as the interpolation error introduced by rescaling, especially up-sampling, is mitigated on the inputs. With branch distillation and backward-compatible training, BTNet transfers discriminative high-resolution information to multiple branches while guaranteeing representation compatibility. Our experiments demonstrate strong performance on face recognition benchmarks, both for multi-resolution identity matching and feature aggregation, with much less computation amount and parameter storage. We establish new state-of-the-art on the challenging QMUL-SurvFace 1: N face identification task.

1 Introduction

Machine learning has advanced tremendously driven by deep learning methods, but is still severely challenged by various data specifications, such as data type, structure, scale and size, etc. For instance, face recognition (FR) is a well-established deep learning task, while the performance degrades dramatically in the testing domain that differs from the training one, influenced by factors of variance like resolution, illumination, occlusion, etc.

Most face recognition methods map each image to a point embedding in the common metric space by deep neural networks (DNNs). The dissimilarity of images can be then calculated using various distance metrics (e.g., cosine similarity, Euclidean distance, etc.) for face recognition tasks.

Recent advancements in margin-based loss (e.g., ArcFace [1], MV-Arc-Softmax [2], CurricularFace [3], etc) enhanced discriminability of the metric space, with small intra-identity distance and large inter-identity distance. However, lack of variation in training data still leads to poor generalizability. Various useful methods are utilized to mitigate this issue. The model adapts to factors of variance by augmenting datasets, whereas the large discrepancy in data distribution could potentially weaken the model's ability to extract discriminative features with the same data scale and model structure (see Section 4.3). Fine-tuning is widely used to transfer large pretrained models to new domains with different data specifications. However, this strategy requires one to store and deploy a separate copy of the backbone parameters for every single new domain, which is expensive and often infeasible.

As known, the resolutions of face images in reality may be far beyond the scope covered by the model. As the small feature maps with a fixed spatial extent (e.g., 7×7) are mapped to an embedding

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

with a predefined dimension (e.g., $128 - d$, $512 - d$, etc.) by a fully connected (fc) layer, input images need to be rescaled to a canonical spatial size (e.g., 112×112) before fed into the network. However, up-sampling low-resolution (LR) images introduces the interpolation error (see Section 3.1), deteriorating the recognizable ones which contain enough clues to identify the subject. Even though super-resolution methods [4–10] are widely used to build faces with good visualization, they inevitably introduce feature information of other identities when reconstructing high-resolution (HR) faces. This may lead to erroneous identity-specific features, which are detrimental to risk-controlled face recognition.

Empirically, we can divide inputs by resolution distribution and learn to operate on them via multiple models to achieve high accuracy and efficiency. However, multi-model fashion cannot be applied directly for cross-resolution recognition as representation compatibility among models need to be guaranteed [11–15].

To improve discriminability while ensure the compatibility of the metric space for multi-resolution face representation, we learn the "unified" representation by a partially-coupled Branch-to-Trunk Network (BTNet). It is composed of multiple independent branch networks (BNets) and a shared trunk network (TNet). A resolution-specific BNet is used for a given image, and the output are implanted as feature maps in the feature pyramid of TNet, at a layer with the same resolution.

Furthermore, we find that multi-resolution training can be beneficial to building a strong and robust TNet, and backward-compatible training (BCT) [11] can improve the representation compatibility during the training process of BTNet. To ameliorate the discriminability of tiny faces, we propose branch distillation in intermediate layers, utilizing information extracted from HR images to help the extraction of discriminative features for resolution-specific branches.

Our method is simple and efficient, which breaks the convention of up-sampling the inputs and serves as a general framework that can be easily implemented by several existing methods due to conceptual simplicity. Meanwhile, BTNet is able to reduce the number of FLOPS by operating the inputs without up-sampling, and per-resolution storage cost by only storing the learned branches and resolution-aware BNs [16], while re-using the copy of the trunk model.

We demonstrate that our method performs comparably in various open-set face recognition tasks (1:1 face verification and 1: N face identification), in both settings of multi-resolution identity matching and feature aggregation, while meaningfully reduces the redundant computation cost and parameter storage. In the challenging QMUL-SurvFace 1: N face identification task [17], we establish new state-of-the-art by outperforming prior models. Furthermore, by avoiding the ill-posed problem (i.e., image up-sampling), our approach also effectively reduces the additional noise and uncertainty of the representation, which plays a key role in reliable risk-controlled face recognition.

2 Related Work

Compatible Representation Learning: The task of compatible representation learning aims at encoding features that are interoperable with the features extracted from other models. Shen et. al. [11] first formulated the problem of backward-compatible learning (BCT) and proposed to utilize the old classifier for compatible feature learning. Since the multi-model fashion benefits representation learning with lower computation, our idea of cross-resolution representation learning can be modeled similar to cross-model compatibility [11–15], as metric space alignment for different resolutions. Our goal is achieved by both compatibility-aware network architecture and training strategy.

Knowledge Distillation and Transfer: The concept of knowledge distillation (KD) was first proposed by Hinton et. al. in [18], which can be summarized as employing a large parameter model (teacher) to supervise the learning of a small parameter model (student). Distillation from intermediate features [19–29] is widely adopted to enhance the effectiveness of knowledge transfer. However, due to the "dark knowledge" hidden in the intermediate layers, additional subtle design is often required to match and rescale intermediate features. Instead, our approach can easily locate the distillation features without rescaling and effectively transfer knowledge from the HR domain to LR branches.

Low Resolution Face Recognition: Its task includes low resolution-to-low resolution (LR-to-LR) matching and low resolution-to-high resolution (LR-to-HR) matching [30]. The work can be divided into two categories [31]: (1) Super-resolution (SR) based methods aim to upscale LR images to construct HR images and use them for feature extraction [4–10]. (2) Projection-based methods aim to extract adequate representations in different domains and project them into a common feature space [32–34]. SR approaches are able to build faces with good visualization, but inevitably introduce feature information of other identities when reconstructing corresponding HR faces, thus introducing noise for identity-specific features. Compared to previous projection methods, our approach directly learns discriminative representations in a common feature space for HR and LR inputs, without additional projection heads for feature transformation.

Pseudo-Siamese Networks: Siamese networks are a coupling architecture based on DNNs, which are widely used for signature verification [35], face verification [36, 37], tracking [38], etc. Pseudo-Siamese networks [39] are decoupled Siamese networks, as the weights of the two branches are not shared, resulting in a more flexible representation way for the two entities. Hughes et. al. in [40] proposed a pseudo-Siamese CNN for identifying corresponding patches in SAR and optical images. Inspired by pseudo-Siamese networks, we propose a resolution-adaptive partially coupled Siamese network architecture, extracting specific-shared features for images with different resolutions.

3 Learning Specific-Shared Feature Transfer

Instead of rescaling the inputs to a canonical size, we build multiple resolution-specific branches (BNets) that are used to map inputs to intermediate features with the same resolution and a resolution-shared trunk (TNet) to map feature maps with different resolutions to a high-dimension embedding. We gain several important properties by doing so: (1) Processing inputs on its original resolution can diminish the inevitably introduced error via up-sampling or information loss via down-sampling, thus preserving the discriminability of visual information with different resolutions. (2) Information streams of different resolutions are encoded uniformly, thus enabling the representation compatibility, which is particularly beneficial to open-set face recognition considering that a compatible metric space is the prerequisite for computing similarity. (3) This also effectively reduce the computation for LR images by supplying computational resources conditioned on the input resolution.

3.1 Up-Sampling Error Analysis

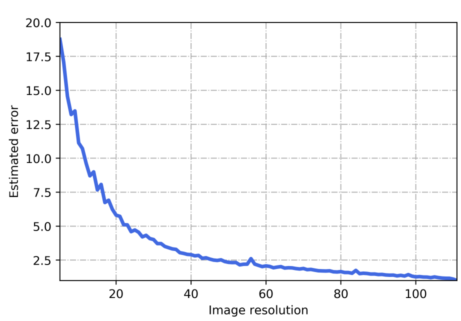

Figure 1: Estimated Error Upperbound. (bilinear interpolation, average value for over 100 images) with the change of image resolution relative to resolution 112.

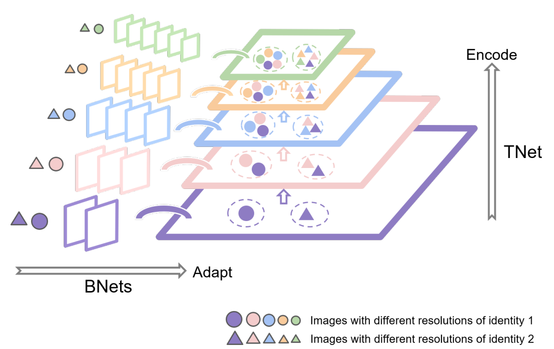

Figure 2: Basic ideas of the proposed BTNet. Images of a certain identity are first projected to the feature maps with the same resolution respectively (Adapt) and then projected to a unified feature representation (Encode). In this figure, feature maps with the same resolution are indicated by outlines in the same color.

Figure 1 illustrates the experimental estimation of interpolation error, whose upper bound increases with the decline of the image resolution (see detailed theoretical derivation in Appendix A.1). Note that the error soars up when the resolution drops below 32 approximately which can be viewed as LR face images, consistent with the tiny-object criterion [41].

The results show that: (1) inputs with a resolution higher than around 32 can be considered in the same HR domain, since the error information introduced by up-sampling via interpolation can be ignored to a certain extent; (2) inputs with a resolution lower than around 32 should be treated as in various LR domains due to the high sensitivity of the resolution to errors.

3.2 Branch-to-Trunk Network

Let X be an input RGB image with a space shape: $X \in \mathbb{R}^{H \times W \times 3}$ where $H \times W$ corresponds to the spatial dimension of the input. For efficient batch training and inference, we predefine a canonical size $S \times S$ (e.g., 112×112 for typical face recognition models like ArcFace [1]).

We build a trunk network $T : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{C_{emb}}$ capable of extracting discriminative information with different resolutions, where C_{emb} is the number of embedding channels. For every resolution r in the candidate set, we formulate a resolution-specific branch, $z_r = B_r(X_r)$, which maps the input image X_r to feature maps with the same resolution and expanded channels $z_r : \mathbb{R}^{r \times r \times 3} \rightarrow \mathbb{R}^{r \times r \times C_r}$. The idea is to learn our branches B to focus on resolution-specific feature transfer independently. Feature maps will then be coupled to the trunk network T in the feature pyramid with the same spatial resolution $r \times r$, allowing for further mapping to the unified presentation space by $T_r : \mathbb{R}^{r \times r \times C_r} \rightarrow \mathbb{R}^{C_{emb}}$.

Here, we follow the idea of "avoiding redundant up-sampling". Our branches B are implemented with same-resolution mapping: i.e., the model preserves the network architecture of T from input to the layer with resolution r and abandons down-sampling operations (e.g., replacing the convolution of stride 2 with stride 1, abandoning the pooling layers, etc.) to keep the same-resolution flow.

We specifically name our specific-shared feature transfer network as Branch-to-Trunk Network, abbreviated as "BTNet". Figure 2 visually summarizes the main ideas of BTNet.

3.3 Training Objectives

We now describe the training objectives. The training of BTNet includes training the trunk network T such that it can produce discriminative and compatible representations for multi-resolution information, and fine-tuning the branch networks B to encourage them to learn resolution-specific feature transfer, so as to improve accuracy without compromising compatibility.

Influence Loss. It is a compatibility-aware classification loss which is implemented by feeding the embeddings of the new model to the classifier of the old model [11]. Since the difficulties of samples vary due to image resolution, we compute CurricularFace [3] as our classification loss, in the form of:

$$L_{cur} = -\log\left(\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j=1, j\neq y_i}^{n} e^{sN(t^{(k)}, \cos(\theta_j))}}\right) \tag{1}$$

$$N(t, \cos\theta_j) = \begin{cases} \cos(\theta_j), & \cos(\theta_{y_i}+m) - \cos(\theta_j) \geq 0 \\ \cos(\theta_j)(t + \cos(\theta_j)), & else \end{cases} \tag{2}$$

$$t^{(k)} = \alpha \sum_i \cos\theta_{y_i} + (1-\alpha)\, t^{(k-1)} \tag{3}$$

which distinguishes both the difficultness of different samples in each stage and relative importance of easy and hard samples during different training stages. Thus, we refine CurricularFace loss as our influence loss:

$$L_{influence} = L_{cur}(\varphi_{bt}, \kappa^*) \tag{4}$$

where φ_{bt} is BTNet backbone (both B_r and T_r), and κ^* is the classifier of the pretrained trunk T.

Branch Distillation Loss. Due to the continuity of the scale change of both the image pyramid and the feature pyramid [42], we can get a qualitative sense of the similarity between images and feature maps with the same resolution (see Figure 3). Furthermore, features extracted from HR images have richer and clearer information than those from LR images [43]. Motivated by these analyses, we utilize an MSE loss to encourage the branch output z_r to be similar to the corresponding feature maps of the pretrained trunk network z_s:

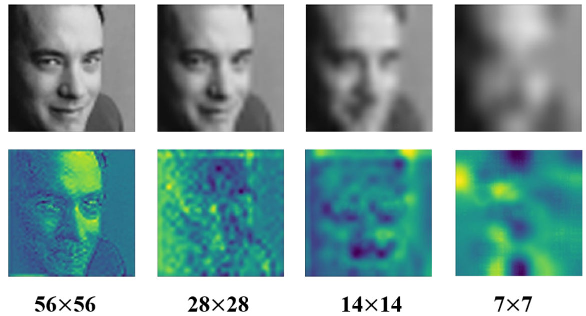

Figure 3: Visual comparison of face image-feature map pairs with different resolutions (resized to a common size here for illustration).

$$L_{branch} = \frac{1}{V}\sum_{v=1}^{V}(z_{r_v} - z_{s_v})^2 \tag{5}$$

where V denotes the batch size.

The whole training objective is a combination of the above objectives:

$$L = L_{influence} + \lambda_{branch} L_{branch} \tag{6}$$

where λ_{branch} is a hyper-parameter to weigh the losses and we set $\lambda_{branch} = 0.5$ in all our experiments.

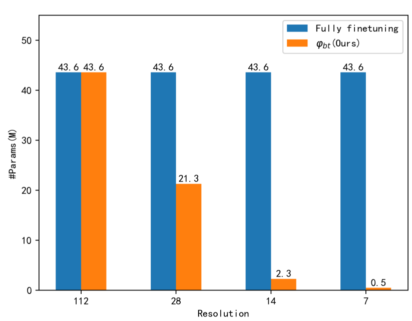

Figure 4: Comparison of # Params (M) between fully finetuning and φ_{bt}.

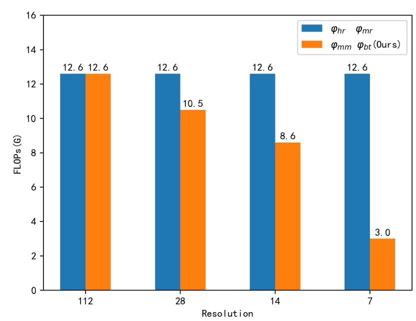

Figure 5: Comparison of FLOPs (G) between baselines and φ_{bt}.

3.4 Storing Branch Networks

An obvious adaptation strategy is fully finetuning of the model on each resolution. However, this strategy requires one to store and deploy a separate copy of the backbone parameters for every resolution, which is an expensive proposition and difficult to expand into more segmented resolution branches. Our BTNet is beneficial in the scenario of multi-resolution face recognition which achieves better parameter/accuracy trade-offs. Since activation statistics including means and variances under different resolutions are incompatible [44], we update and store Batch Normalization (BN) [45] parameters in all layers of B_r and T_r for each resolution, whose amount is negligible. Apart from this, we only need to store the learned branches and re-use the original copy of the pretrained trunk model, significantly reducing the storage cost. Figure 4 shows that BTNet requires only $1.1\% \sim 48.9\%$ of all the parameters compared to fully updating all the parameters of TNet.

4 Experiments

To validate BTNet on face recognition tasks in open universe, we perform 1:1 verification and $1:N$ identification tasks in two different settings, including (a) multi-resolution identity matching, and

(b) multi-resolution feature aggregation. For 1:1 verification, a pair of templates are provided and the model is to decide whether they belong to the same identity or not. For 1:N identification, a set of gallery images are first mapped onto their embedding vectors (indexing) and the embeddings of query images are extracted to perform search against indexed gallery.

4.1 Implementation Details

Datasets. We use MS1Mv3 [46] for training face embedding models. The MS1Mv3 dataset contains 5,179,510 images of 93,431 celebrities. According to the test setting, different test datasets are used.

·Multi-Resolution Identity Matching. We try on six widely adopted face verification benchmarks: LFW [47], CFP-FF [48], CFP-FP [48], AgeDB-30 [49], CALFW [50], and CPLFW [51], while the large-scale surveillance face dataset QMUL-SurvFace [17] is used for 1:N face identification, which contains native LR surveillance faces across wide space and time. The spatial resolution for QMUL-SurvFace ranges from 6/5 to 124/106 in height/width with an average of 24/20.

·Multi-Resolution Feature Aggregation. We adopt a top challenging benchmark IJB-C [52], which has around 130k images from 3,531 identities, for two standard testing protocols: $1:1$ verification and 1:N identification.

Training. All the models are trained on four RTX 2080 Tis with batch size 128 by stochastic gradient descent. For TNet, we train for 25 epochs, with learning rate initialized at 0.2 with 2 warm-up epochs and decaying as a quadratic polynomial. We augment training samples by random horizonal flipping and multi-resolution training. For BNets, we initialize the learning rate by 0.02 without warm-up epochs. The training all stops at the $10th$ epoch for a fair comparison. The recommended hyper-parameters are used for classification loss from the original paper (e.g., $m = 0.5, s = 64$ for ArcFace [1], and $\alpha = 0.99, t^0 = 0$ for CurricularFace [3]). Only horizonal flipping is used as augmentation when training BNets.

Baselines. In our experiment, several baselines are used to validate BTNet in learning discriminative and compatible representations for multi-resolution face recognition.

·High-Resolution Trained φ_{hr}. Naive baseline trained with HR data.

·Independently Trained φ_{mm}. Multi-model fashion: is it possible to achieve better results if we train a specific model for each resolution independently? Specifically, we train φ_r for data with resolution r and denote the multi-model collections as φ_{mm}.

·Multi-Resolution Trained φ_{mr}. Trained with multi-resolution data which adapts to resolution-variance. Specifically, each image is randomly down-sampled to a size in the candidate set $\{\frac{112}{2^i} \times \frac{112}{2^i} | i = 0, 1, 2, 3, 4\}$ with equal probability of being chosen, and then up-sampled back to 112×112.

Instantiation of Network Architecture. The BTNet and baselines are implemented with ResNet50 [53], and they could be extended easily with other implementations. Dubbed as φ_{bt}, the detailed instantiation of BTNet based on ResNet50 is illustrated in Appendix A.2.

4.2 Evaluation Metrics

On the benchmarks for face verification, we use 1:1 verification accuracy as the basic metrics. The rank-20 true positive identification rates (TPIR20) at varying false positive identification rates (FPIR) and AUC are used to report the identification results on QMUL-SurvFace. The evaluation metrics for IJB-C 1:1 verification protocol are true acceptance rates (TAR) at different false acceptance rate (FAR). For 1:N identification, the basic evaluation metrics are the true positive identification rates (TPIR) at different false positive identification rates (FPIR).

For better evaluation, we define another two metrics to assess the relative performance gain similar to [11, 14].

Table 1: Comparison of different methods on six face verification benchmarks.
"Acc." denotes average 1:1 verification accuracy.

(a) Cross-resolution identity matching.

	112&7		112&14		112&28	
	Acc.	Gain	Acc.	Gain	Acc.	Gain
φ_{hr}	57.75	-	81.02	-	95.90	-
φ_{mm}	50.58	-0.89	49.90	-4.82	50.03	-305.80
φ_{mr}	65.85	+1.00	87.47	+1.00	96.05	+1.00
φ_{bt}(Ours)	**86.10**	**+3.50**	**94.08**	**+2.02**	**96.65**	**+5.00**

(b) Same-resolution identity matching.

	7&7		14&14		28&28		112&112	
	Acc.	Gain	Acc.	Gain	Acc.	Gain	Acc.	Gain
φ_{hr}	60.70	-	73.88	-	93.58	-	**97.68**	-
φ_{mm}	62.57	+1.00	78.00	+1.00	94.68	+1.00	**97.68**	-
φ_{mr}	61.02	+0.17	80.32	+1.56	95.12	+1.40	97.25	-
φ_{bt}(Ours)	**77.78**	**+9.13**	**90.90**	**+4.13**	**96.27**	**+2.45**	97.25	-

Cross-Resolution Gain. With the purpose towards the cross-resolution compatible representations, we define the performance gain as follows:

$$Gain_{r_1\&r_2}(\varphi) = \frac{M_{r_1\&r_2}(\varphi) - M_{r_1\&r_2}(\varphi_{hr})}{|M_{r_1\&r_2}(\varphi_{mr}) - M_{r_1\&r_2}(\varphi_{hr})|} \tag{7}$$

Here $M_{r_1\&r_2}(\cdot)$ are metrics when the resolutions of the image/template pair are $r_1 \times r_1$ and $r_2 \times r_2$ ($r_1 \neq r_2$), respectively. φ_{mr} shares the same architecture with φ_{hr} while is trained on multi-resolution images and thus serves as the baseline of cross-resolution gain.

Same-Resolution Gain. For the scenario of multi-resolution face recognition, the performance of same-resolution verification/identification is also vital besides cross-resolution one. Therefore, we report the relative performance improvement from base model φ_{hr} in the scenario of same-resolution.

$$Gain_{r\&r}(\varphi) = \frac{M_{r\&r}(\varphi) - M_{r\&r}(\varphi_{hr})}{|M_{r\&r}(\varphi_r) - M_{r\&r}(\varphi_{hr})|} \tag{8}$$

Here $M_{r\&r}(\cdot)$ are metrics when the resolutions of the image/template pair are both $r \times r$. φ_r is a model of the set $\{\varphi_{mm} = \varphi_r | r = 7, 14, 28\}$ trained on images with resolution $r \times r$ without considering cross-resolution representation compatibility, which serves as the baseline of same-resolution gain on resolution r. Note that for both metrics we add the absolute symbol to the denominator as they can be negative in some test settings (detailed in Section 4.3 and 4.4).

4.3 Multi-Resolution Identity Matching

We now conduct experiments on the proposed BTNet framework for multi-resolution identity matching. Two different settings are included : (1) same-resolution matching, and (2) cross-resolution matching. Table 1 compares the average performance on popular benchmarks for φ_{hr}, φ_{mm}, φ_{mr}, φ_{bt}. The experimental results on each dataset are detailed in Appendix A.5.

When directly applied to test data with the resolution lower than training data, φ_{hr} suffers a severe performance degradation. Up-sampling images via interpolation can increase the amount of data but not the amount of information, only to improve the detailed part of the image and the spatial resolution (size) [64]. Moreover, it also brings various noise and artificial processing traces [65]. Up-sampling images via interpolation-typically bilinear interpolation or bicubic interpolation of 4x4 pixel neighborhoods, essentially a function approximation method, is bound to introduce error information (detailed in Appendix A.1), thus potentially confusing identity information, which is especially crucial for LR images with limited details. We are able to observe improvement of φ_{mm} in same-resolution matching but its cross-resolution gain is negative with approximately 50% accuracy. Unsurprisingly, independently trained φ_r is unaware of representation compatibility, and thus does not naturally suitable for cross-resolution recognition. The results show that φ_{mr} improved both cross-resolution and same-resolution accuracy by a large margin, as it learns to adapt to resolution variance and maintain discriminability of multi-resolution inputs. Note that the model size and training data scale stay the same, while only the resolution distribution of the data changes for φ_{mr}, and thus there is a marginal accuracy drop in the setting of 112&112 matching. Comparably, φ_{bt} substantially outperforms all baselines with 2.02 ~5.00 cross-resolution gain and 2.45~9.13 same-resolution gain. Importantly, due to the multi-resolution branches, our approach has a cost same with φ_{mm}, significantly lower than φ_{hr} and φ_{mr} (see Figure 5).

Table 2: Performance of face identification on QMUL-SurvFace. Most compared results are cited from [17, 54] except BTNet.

	TPIR20(%)@FPIR				
	AUC	0.3	0.2	0.1	0.01
VGG-Face [55]	14.0	5.1	2.6	0.8	0.1
DeepID2 [56]	20.8	12.8	8.1	3.4	0.8
FaceNet [57]	19.8	12.7	8.1	4.3	1.0
SphereFace [58]	28.1	21.3	15.7	8.3	1.0
SRCNN [59]	27.0	20.0	14.9	6.2	0.6
FSRCNN [60]	27.3	20.0	14.4	6.1	0.7
VDSR [61]	27.3	20.1	14.5	6.1	0.8
DRRN [62]	27.5	20.3	14.9	6.3	0.6
LapSRN [63]	27.4	20.2	14.7	6.3	0.7
ArcFace [1]	25.3	18.7	15.1	10.1	2.0
RAN [54]	32.3	26.5	21.6	14.9	**3.8**
BTNet (avg.+floor)	32.6	27.9	23.4	16.5	1.4
BTNet (avg.+near)	34.6	30.3	25.7	18.9	1.5
BTNet (avg.+ceil)	**35.4**	31.1	26.8	20.3	2.2
BTNet (min+floor)	32.3	27.6	23.2	16.1	1.4
BTNet (min+near)	34.0	29.6	25.0	18.0	1.4
BTNet (min+ceil)	35.3	31.0	26.6	19.9	2.0
BTNet (max+floor)	33.6	29.1	24.5	17.6	1.3
BTNet (max+near)	35.2	31.0	26.4	19.6	1.7
BTNet (max+ceil)	**35.4**	**31.2**	**26.9**	**20.6**	2.5

For inference on inputs with resolutions not strictly matched to the branch, we validate three selection strategies based on three resolution indicators (see Figure 6). Table 2 compares BTNet against the state-of-the-arts models on QMUL-SurvFace 1:N identification benchmark. We are able to observe that our proposed approach extends the state-of-the-arts while being more computationally efficient. We believe the performance of BTNet (max + ceil) is the highest that have been reported so far, and we believe it is meaningful with the increased focus on unconstrained surveillance applications.

4.4 Multi-Resolution Feature Aggregation

Multi-resolution feature aggregation is common in set-based recognition tasks where the model needs to determine the similarity of sets (templates), instead of images. Each set could contain images of the same identity with different resolutions. In our experiment, we rescale the original and flipped images in each set to different resolutions and aggregate their features into a representation of the template. Detailed experimental results can be seen in Appendix A.5.

Table 3 (a) compares the cross-resolution results of TAR@FAR=10^{-4} for 1:1 verification. The cross-resolution features are ensured to be mapped to the same vector space where the aggregation is conducted for φ_{hr} and φ_{mr}, but we can observe that φ_{hr} performs much better than φ_{mr}. One possible reason is that φ_{hr} has outstanding discriminability to extract HR features, while LR features may not overly deteriorate the HR information. This phenomenon also suggests that φ_{mr} sacrifices its discriminability in exchange for the adaptability for resolution-variance. We can see φ_{bt} is comparable with φ_{hr}, demonstrating the discriminative power of BTNet for aggregating multi-resolution features.

Table 3 (b) compares the same-resolution results of TAR@FAR=10^{-4} for 1:1 verification. When HR information is removed from the template representation (i.e., test settings 7&7, $14\&14$, $28\&28$), φ_{hr} suffers from performance degradation as well, as the informative embedding cannot catch the lost details of the LR images [54]. Both φ_{mm} and φ_{mr} improve with a limited same-resolution gain, while φ_{bt} surpasses the baselines by a large margin while also reducing the compute.

In Table 4 we show the results of TPIR@FPIR=10^{-1} for 1:N identification protocol. Similar to our results for 1:1 verification, we are able to observe that φ_{bt} is comparable or even better than φ_{hr} with

Table 3: Comparison of different methods on the IJB-C dataset 1:1 face verification task. "TAR" denotes TAR (%@FAR=1e-4).

(a) Cross-resolution feature aggregation.

	112&7		112&14		112&28	
	TAR	Gain	TAR	Gain	TAR	Gain
φ_{hr}	**88.89**	-	92.40	-	**95.62**	-
φ_{mm}	74.54	-0.56	93.52	+1.33	95.42	-0.69
φ_{mr}	63.11	-1.00	91.56	-1.00	95.33	-1.00
φ_{bt}(Ours)	88.17	-0.03	**93.97**	**+1.87**	**95.62**	**+0.00**

(b) Same-resolution feature aggregation.

7&7		14&14		28&28		112&112	
TAR	Gain	TAR	Gain	TAR	Gain	TAR	Gain
4.83	-	33.74	-	89.65	-	**96.40**	-
4.83	+ 0.00	29.26	-1.00	92.58	+1.00	**96.40**	-
4.48	-	40.51	+1.51	92.81	+1.08	96.06	-
35.47	-	**82.08**	**+10.79**	**94.50**	**+1.66**	96.06	-

Table 4: Comparison of different methods on the IJB-C dataset 1: N face identification task. "TPIR" denotes TPIR (%@FPIR=0.1).

(a) Cross-resolution feature aggregation.

	112&7		112&14		112&28	
	TPIR	Gain	TPIR	Gain	TPIR	Gain
φ_{hr}	**85.60**	-	90.11	-	94.27	-
φ_{mm}	69.70	-0.55	91.73	+1.53	94.13	-0.33
φ_{mr}	56.64	-1.00	89.05	-1.00	93.84	-1.00
φ_{bt}(Ours)	83.93	-0.06	**91.87**	**+1.66**	**94.33**	**+0.14**

(b) Same-resolution feature aggregation.

7&7		14&14		28&28		112&112	
TPIR	Gain	TPIR	Gain	TPIR	Gain	TPIR	Gain
3.12	-	26.37	-	86.06	-	**95.57**	-
3.24	+1.00	21.84	-1.00	89.76	+1.00	95.57	-
3.25	+1.08	37.58	+2.47	91.02	+1.34	94.85	-
27.70	**+204.83**	**76.65**	**+11.10**	**92.89**	**+1.85**	94.85	-

HR information involved and can preserve superior discriminability with limited LR information, while also being more computationally efficient.

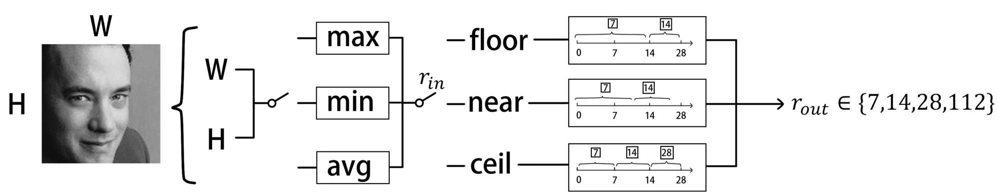

Figure 6: Branch selection process. Max/min/average is used on (W, H) to obtain a resolution indicator for further allocation (floor/near/ceil) to a certain branch.

5 Discussion and Conclusion

This paper works on the problem of multi-resolution face recognition, and provides a new scheme to operate images conditioned on its input resolution without large span rescaling. The error introduced by up-sampling via interpolation is investigated and analyzed. Decoupled as branches for discriminative representation learning and coupled as the trunk for compatible representation learning, our Branch-to-Trunk Network (BTNet) achieves significant improvements on multi-resolution face verification and identification tasks. Besides, the superiority of BTNet in reducing computational cost and parameter storage cost is also demonstrated. It is worth noting that our approach is easy to expand to recognition tasks for other classes of objects and has the potential to serve as a general network architecture for multi-resolution visual recognition.

Limitations and Future Work. The dislocation between the underlying optical resolution of native face images and that of a certain branch may limit the power of the model, which may be improved by selecting the optimal processing branch for the input in combination with the image quality, rather than by image size alone. The optimal branch selection strategy is not fully investigated though we have provided an intuitive way to select the branch for inputs (see Figure 6). Importantly, based on the unified multi-resolution metric space, the underlying resolution of the inputs (integrated spatial resolution with quality assessment) can be utilized to provide the reliability of the representation and contribute to risk-controlled face recognition. They will be our future research directions.

References

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019.

[2] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12241–12248. AAAI Press, 2020.

[3] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5900–5909. Computer Vision Foundation / IEEE, 2020.

[4] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 614–630, Cham, 2016. Springer International Publishing.

[5] Klemen Grm, Walter J. Scheirer, and Vitomir Struc. Face hallucination using cascaded super-resolution and identity priors. *IEEE Trans. Image Process.*, 29:2150–2165, 2020.

[6] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S. Huang. Studying very low resolution recognition using deep networks. *CoRR*, abs/1601.04153, 2016.

[7] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part III*, volume 11363 of *Lecture Notes in Computer Science*, pages 605–621. Springer, 2018.

[8] Xi Yin, Ying Tai, Yuge Huang, and Xiaoming Liu. FAN: feature adaptation network for surveillance face recognition and normalization. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi, editors, *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part II*, volume 12623 of *Lecture Notes in Computer Science*, pages 301–319. Springer, 2020.

[9] Maneet Singh, Shruti Nagpal, Richa Singh, and Mayank Vatsa. Dual directed capsule network for very low resolution image recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 340–349. IEEE, 2019.

[10] Aashish Rai, Vishal M. Chudasama, Kishor P. Upla, Kiran B. Raja, Raghavendra Ramachandra, and Christoph Busch. Comsupresnet: A compact super-resolution network for low-resolution face images. In *8th International Workshop on Biometrics and Forensics, IWBF 2020, Porto, Portugal, April 29-30, 2020*, pages 1–6. IEEE, 2020.

[11] Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6367–6376. Computer Vision Foundation / IEEE, 2020.

[12] Mateusz Budnik and Yannis Avrithis. Asymmetric metric learning for knowledge transfer. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8228–8238. Computer Vision Foundation / IEEE, 2021.

[13] Chien-Yi Wang, Ya-Liang Chang, Shang-Ta Yang, Dong Chen, and Shang-Hong Lai. Unified representation learning for cross model compatibility. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[14] Qiang Meng, Chixiang Zhang, Xiaoqiang Xu, and Feng Zhou. Learning compatible embeddings. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9919–9928. IEEE, 2021.

[15] Rahul Duggal, Hao Zhou, Shuo Yang, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Compatibility-aware heterogeneous visual search. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10723–10732. Computer Vision Foundation / IEEE, 2021.

[16] Mingjian Zhu, Kai Han, Enhua Wu, Qiulin Zhang, Ying Nie, Zhenzhong Lan, and Yunhe Wang. Dynamic resolution network. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27319–27330, 2021.

[17] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Surveillance face recognition challenge. *CoRR*, abs/1804.09691, 2018.

[18] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[19] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1921–1930. IEEE, 2019.

[20] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *CoRR*, abs/1707.01219, 2017.

[21] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3967–3976. Computer Vision Foundation / IEEE, 2019.

[22] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[23] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1365–1374. IEEE, 2019.

[24] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016.

[25] Junho Yim, Donggyu Joo, Ji-Hoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7130–7138. IEEE Computer Society, 2017.

[26] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *CoRR*, abs/1910.10699, 2019.

[27] Baoyun Peng, Xiao Jin, Dongsheng Li, Shunfeng Zhou, Yichao Wu, Jiaheng Liu, Zhaoning Zhang, and Yu Liu. Correlation congruence for knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5006–5015. IEEE, 2019.

[28] Jangho Kim, Seonguk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2765–2774, 2018.

[29] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3779–3787. AAAI Press, 2019.

[30] Yoanna Martínez-Díaz, Heydi Méndez-Vázquez, Luis S. Luevano, Leonardo Chang, and Miguel González-Mendoza. Lightweight low-resolution face recognition for surveillance applications. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 5421–5428. IEEE, 2020.

[31] Luis S. Luevano, Leonardo Chang, Heydi Méndez-Vázquez, Yoanna Martínez-Díaz, and Miguel González-Mendoza. A study on the performance of unconstrained very low resolution face recognition: Analyzing current trends and new research directions. *IEEE Access*, 9:75470–75493, 2021.

[32] Ze Lu, Xudong Jiang, and Alex C. Kot. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Process. Lett.*, 25(4):526–530, 2018.

[33] Sivaram Prasad Mudunuri, Soubhik Sanyal, and Soma Biswas. Genlr-net: Deep framework for very low resolution face and object recognition with generalization to unseen categories. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 489–498. Computer Vision Foundation / IEEE Computer Society, 2018.

[34] Juan Zha and Hongyang Chao. TCN: transferable coupled network for cross-resolution face recognition*. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 3302–3306. IEEE, 2019.

[35] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a siamese time delay neural network. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pages 737–744. Morgan Kaufmann, 1993.

[36] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1701–1708. IEEE Computer Society, 2014.

[37] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 539–546. IEEE Computer Society, 2005.

[38] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. In Gang Hua and Hervé Jégou, editors, *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, volume 9914 of *Lecture Notes in Computer Science*, pages 850–865, 2016.

[39] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4353–4361. IEEE Computer Society, 2015.

[40] Lloyd H. Hughes, Michael Schmitt, Lichao Mou, Yuanyuan Wang, and Xiao Xiang Zhu. Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN. *IEEE Geosci. Remote. Sens. Lett.*, 15(5):784–788, 2018.

[41] Antonio Torralba, Robert Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, 2008.

[42] Tony Lindeberg. Scale-space theory: a basic tool for analyzing structures at different scales. *Journal of Applied Statistics*, 21(1-2):225–270, 1994.

[43] Yui Man Lui, David Bolme, Bruce A. Draper, J. Ross Beveridge, Geoff Givens, and P. Jonathon Phillips. A meta-analysis of face recognition covariates. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–8, 2009.

[44] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8250–8260, 2019.

[45] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015.

[46] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 2638–2646. IEEE, 2019.

[47] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 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*, 2008.

[48] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

[49] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1997–2005. IEEE Computer Society, 2017.

[50] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017.

[51] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5:7, 2018.

[52] Brianna Maze, Jocelyn C. Adams, James A. Duncan, Nathan D. Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA janus benchmark - C: face dataset and protocol. In *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*, pages 158–165. IEEE, 2018.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[54] Han Fang, Weihong Deng, Yaoyao Zhong, and Jiani Hu. Generate to adapt: Resolution adaption network for surveillance face recognition. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV*, volume 12360 of *Lecture Notes in Computer Science*, pages 741–758. Springer, 2020.

[55] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Xianghua Xie, Mark W. Jones, and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pages 41.1–41.12. BMVA Press, 2015.

[56] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1988–1996, 2014.

[57] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society, 2015.

[58] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6738–6746. IEEE Computer Society, 2017.

[59] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, volume 8692 of *Lecture Notes in Computer Science*, pages 184–199. Springer, 2014.

[60] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 391–407. Springer, 2016.

[61] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1646–1654. IEEE Computer Society, 2016.

[62] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2790–2798. IEEE Computer Society, 2017.

[63] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5835–5843. IEEE Computer Society, 2017.

[64] Z. G. Liu and D. Z. Liu. Reappraising about image magnification methods based on wavelet transformation. *Journal of Image and Graphics*, 2003.

[65] Wan-Chi Siu and Kwok-Wai Hung. Review of image interpolation and super-resolution. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2012, Hollywood, CA, USA, December 3-6, 2012*, pages 1–10. IEEE, 2012.

[66] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8681–8691. Computer Vision Foundation / IEEE, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.The dislocation between the underlying optical resolution of native face images and that of a certain branch may limit the power of the model, which may be improved by selecting the optimal processing branch for the input in combination with the image quality, rather than by image size alone. The optimal branch selection strategy is not fully investigated though we have provided an intuitive way to select the branch for inputs.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We study a general framework for multi-resolution face recognition. Our method is not for specific applications, which does not directly involve societal issues.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Appendix A.1
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.1
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We follow the common practice in previous works, where they didn't report the error bars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See our implementation details in Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]